THE UNIVERSITY
of EDINBURGH

THE REGULATORY MECHANISMS AND BIOLOGICAL IMPLICATIONS OF PROTEIN COMPLEX ASSEMBLY

Jonathan N. Wells

Doctor of Philosophy
The University of Edinburgh
2017

DECLARATION

This thesis presents my own work, and has not been submitted for any other degree or professional qualification. Wherever results were obtained in collaboration with others, I have clearly stated it in the text. Any information derived from the published work of others has been cited in the text, and a complete list of references can be found in the bibliography. Published papers arising from the work described in this thesis can be found in the appendices.

– Jonathan Wells, 2017

To my parents, Jane and Nick

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my PhD supervisor, Joe Marsh. Joe has been a fantastic mentor, providing me with invaluable research opportunities, sound advice and the freedom to pursue my own ideas, all the while with patience and good humour. I must also thank my colleagues in the Marsh lab: Therese, György and Marcin, for numerous useful and enjoyable discussions. Beyond Edinburgh, I have had the good fortune of being able to collaborate with other scientists from around the world, and whilst there are too many to name individually, it goes without saying that this thesis would not have been possible without them.

Even jobs as good as doing science for a living get tiresome sometimes, and in these moments I am very grateful to have the friends I do; no problem is so great that it can't be lessened by laughter and good company, both of which they provide in abundance. Finally, I would like to thank my parents for all their support over the last few years, and my brothers for being such great brothers.

ABSTRACT

Every living organism possesses a genome that contains within it a unique set of genes, a substantial number of which encode proteins. Over the last 20 years, it has become apparent that organismal complexity arises not from the specific complement of genes per se, but rather from interactions between the gene products - in particular, interactions between proteins. As an inevitable consequence of the crowded cellular interior, most protein-protein interactions are fleeting. However, many are significantly more long-lived and result in stable protein complexes, in which the constituent subunits are obligately dependent on their binding partners. Despite the abundance of protein complexes and their critical importance to the cell, we currently have an incomplete understanding of the mechanisms by which the cell ensures their correct assembly.

In the chapters that follow, I have attempted to improve our understanding of the regulatory systems underlying assembly of protein complexes, and the way in which assembly as a whole affects the behaviour of the cell. The thesis opens with an extended literature review covering the currently available methods for characterising protein complexes. After this introduction, chapters 2-4 are concerned with regulatory mechanisms and biological implications common to the assembly of all protein complexes. Chapter 5 diverges from this work, and describes a family of evolutionarily related proteins that regulate the behaviour of condensins and cohesins.

Bacterial and archaeal genomes contain far less non-coding DNA than eukaryotes, and coding genes are often packaged into discrete units known as operons. The proteins encoded within operons are usually functionally related, either through participation in metabolic pathways or as subunits of heteromeric protein complexes. Since protein complexes assemble via ordered pathways, we reasoned that there might be a signature of assembly order present in operons, the genes of which are translated in sequential order. By comparing computationally predicted assembly pathways with gene order in operons, we demonstrated this to be the case for the large majority of operon-encoded complexes. Within operons, gene order follows assembly order, and adjacent genes are substantially more likely to share a physical interface than those further apart. This work demonstrates that efficient assembly of complexes is of sufficient importance as to have placed major constraints on the evolution of operon gene order.

Following this study of bacterial operons, I present results from research investigating how patterns of protein degradation in eukaryotes are influenced by the formation of protein complexes. This showed that, whilst most proteins display exponential degradation kinetics, a sizeable minority deviate considerably from this pattern, instead being more consistent with a two-step degradation process. These proteins are predominantly members of heteromeric complexes, and their two-step decay profiles can be explained using a model under which bound and unbound subunits are de-

graded at different rates. Within individual complexes, we find that non-exponentially decaying proteins tend to form larger interfaces, assemble earlier, and show a higher degree of coexpression, consistent with the idea that bound subunits are degraded at a slower rate than unbound or peripheral subunits.

This model also explains the behaviour of proteins in aneuploid cells where one or more chromosomes have been duplicated. In general, protein abundance scales with gene copy number, so that the immediate effect of duplicating a chromosome is to double the abundance of the proteins encoded on it. However, previous analyses of mass spectrometry data, as well as my own, have shown that the abundance of many proteins on duplicated chromosomes is significantly attenuated compared to what one would expect. These proteins, like those with non-exponential degradation patterns, are very often members of larger complexes. Since the overall concentration of a protein complex is constrained by that of its least abundant members, duplicating a single subunit will predominantly increase the unbound, unstable fraction of that subunit. The results from this work strongly suggest that the apparent attenuation of many proteins observed in aneuploid cells is indeed a consequence of the failure of these proteins to assemble into complexes.

Finally, I present a study concerning an important, universally conserved family of protein complexes, namely the SMC-kleisins. Two members of this family, condensin and cohesin, are responsible for two hallmarks of eukaryotic chromatin organisation: the formation of condensed, linear chromosomes, and sister chromatid cohesion during cell division. Unlike other SMC-kleisins, condensin and cohesin possess a number of regulators containing HEAT repeats. By developing a computational pipeline for searching and clustering paralogous repeat proteins, I was able to demonstrate that these regulators form a distinct sub-family within the larger class of HEAT repeat proteins. Furthermore, these regulators arose very early in eukaryotic history, hinting at a possible role in the origin of modern condensins and cohesins.

LAY SUMMARY

All cells are made up of a complex mixture of biological macromolecules, including carbohydrates, lipids, and proteins. Proteins, the subject of this thesis, are tiny, vibrating strings of amino acids with a strictly defined sequence and three-dimensional structure. Every cell in your body, of which there are some 50 trillion, contains further trillions of proteins that, collectively, are responsible for carrying out virtually every biological process you can imagine, from the moment you are born, to the moment you die.

However, although each protein is present in many copies in the cell, the full set of unique protein species is comparatively small. What is more, the number of protein-coding genes that an organism has bears almost no relationship to the perceived complexity of that organism. For a humbling illustration of this, consider the fact that your genome - that of a human - contains approximately 20,000 genes, whereas the pufferfish contains closer to 50,000 and even the lowly banana has more than 36,000. What is it then about this collection of genes that allows us to contemplate the difference between our selves and a banana, whilst the banana just lies there, fruitily?

Part of the explanation stems from the fact that proteins interact extensively with one another. Across the entire proteome (the collection of all proteins present in a cell), a substantial fraction of proteins form stable complexes. Haemoglobin, for example, is made up of two alpha and two beta globin subunits. Without the tendency of proteins such as the globins to interact, the level of complexity that we see across the spectrum of life - even in the simplest microorganisms - would never be possible. However, we know very much less about these protein complexes than we do about their constituent subunits. In particular, we only have a basic understanding of how the cell regulates their assembly, ensuring that proteins are produced at the right time and in the right place.

In this work, I have attempted to explain why protein complexes matter, looking at some of the ways in which the cell enables complexes to assemble, and what the biological implications of this process are. For example, from studying the organisation of bacterial genes, it is clear that the order in which their genes are encoded closely matches the order in which the resulting protein complexes assemble. This implies that these creatures must be under strong evolutionary pressure to assemble protein complexes quickly and efficiently.

Concerning human biology, the later chapters describe a model of protein complex behaviour that explains some of the features that we see in aneuploid cells - that is, cells which have an abnormal number of chromosomes. This is a state familiar to many of us as Down's syndrome, in which people affected have an extra copy of chromosome 21. In closing, the work presented here, supported by that of many others, demonstrates the fundamental importance of protein complexes to life, and I hope goes some way to deepening our understanding of their behaviour within cells.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

2D Two-dimensional

3D Three-dimensional

AIC Akaike information criterion

AMBER Assisted Model Building and Energy Refinement

AP Affinity-purification

CAPRI Critical Assessment of PRediction of Interactions

CASP Critical assessment of protein structure prediction

CHARMM Chemistry at HARvard Molecular Mechanics

CNV Copy number variant

DQE Detective quantum efficiency

ED Exponential degradation

EGTA Ethylene glycol tetraacetic acid

EM Electron microscopy

EPR Electron paramagnetic resonance

ESI Electrospray ionisation

eQTL Expression quantitative trait loci

GO Gene ontology

GTEx Genotype-tissue expression

HEAT Huntingtin, elongation factor 3 (EF3), protein phosphatase 2A (PP2A), target of rapamycin (TOR1)

iBAQ Intensity based absolute quantification

IR Isomorphous replacement

iTRAQ Isobaric tag for relative and absolute quantitation

LC Liquid chromatography

LECA Last eukaryotic common ancestor

LFQ Label-free quantification

MAD Multiple wavelength anomalous diffraction

MALDI Matrix-assisted laser desorption/ionisation

MAPS Monolithic active pixel sensors

MCL Markov cluster algorithm

MD Molecular dynamics

MS Mass spectrometry

NED Non-exponential degradation

NMR Nuclear magnetic resonance

PALM Photoactivated localisation microscopy

PDB Protein Data Bank

RSS Residual sum of squares

SASE Self-amplified spontaneous emission

SILAC Stable isotope labelling and culturing

SMC Structural maintenance of chromosomes

SMLM Single molecule localisation microscopy

STED Stimulated emission depletion

STORM Stochastic optical reconstruction microscopy

TAP Tandem affinity purification

TBM Template-based modelling

TEV Tobacco etch virus

TMV Tobacco mosaic virus

TROSY Transverse relaxation-optimised spectroscopy

Y2H Yeast-2-Hybrid

XL Cross-linking

NOTES ON THE USE OF PUBLISHED MATERIAL

This thesis is predominantly based on work that has already been published. I present here a brief description of how each chapter relates to these published articles, and the extent of my involvement in each case. In all cases, chapters were rewritten and modified to varying degrees, using the published articles as guides to structure where applicable. To make my contribution to each project explicit, throughout this thesis I have used 'I' to refer to analyses carried out solely by myself, and 'we' for those that were carried out in close collaboration with others. Where figures have been adapted from published material, notes indicating this have been made in the legends. Any papers for which I am listed as an author that were released whilst working towards this thesis can be found in appendix A.2.

Chapter 2 is derived from:

> Wells, J. N., Bergendahl, L. T. & Marsh, J. A. Operon Gene Order Is Optimized for Ordered Protein Complex Assembly. *Cell Reports* **14,** 679–685. ISSN: 22111247 (Feb. 2016)

This paper has been rewritten, but is largely unchanged in terms of its content and general layout, with the exception of some supplementary figures that have been moved to the main body of text.

Chapter 3 is based on and heavily adapted from:

> McShane, E., Sin, C., Zauber, H., Wells, J. N., Donnelly, N., Wang, X., Hou, J., Chen, W., Storchova, Z., Marsh, J. A., Valleriani, A. & Selbach, M. Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. *Cell* **167,** 803–815. ISSN: 00928674 (Oct. 2016)

This was a large collaboration lead primarily by E. McShane and M. Selbach. A detailed description of author contributions can be found in the paper; my own contribution (with J. Marsh) was in the design, implementation and interpretation of dry experiments investigating the relationship between protein complex assembly and the observed patterns of protein degradation. It should be noted that numerous analyses are included here which are not present in the published version.

Finally, chapter 5 is an extended version of the following article:

Wells, J. N., Gligoris, T. G., Nasmyth, K. A. & Marsh, J. A. Evolution of condensin and cohesin complexes driven by replacement of Kite by Hawk proteins. *Current Biology* **27,** R17–R18. ISSN: 09609822 (Jan. 2017)

T. Gligoris initially suggested this project, experiments were designed by J. Marsh and myself, and analyses were carried out by myself. Interpretation of the data was predominantly my own, with contributions from all other authors.

I do things like get in a taxi and say,
"The library, and step on it."

– David Foster Wallace

1 INTRODUCTION

1.1 WHAT ARE PROTEIN COMPLEXES?

Earth's biosphere is built around an ancient, universally conserved metabolic network, which channels solar and chemical energy into thermal energy wherever life exists, collectively accounting for a vast energy flux across the face of the planet. The primary catalysts of this network are proteins – large biological macromolecules formed from chains of amino acids. In addition to catalysis, proteins are involved in almost all other biological processes, making them integral to every cell on Earth. Without proteins and the interactions they make with their environment, life would be limited to simple, autocatalytic reactions occurring at the interfaces between water, rocks, and the atmosphere.

Proteins are complex molecules; each one consists of tens to thousands of amino acids linked by covalent peptide bonds between their terminal amine and carboxyl groups. There are 20 amino acids widely used by biological organisms, and thus an N residue sequence can be composed in 20^N different ways - for a typical 200 residue protein this is 1.6×10^{260} possible sequence combinations. In addition to its sequence, the function of a protein is critically dependent on its three-dimensional (3D) structure, as defined by bond angles between pairs of residues. As a result of this sequence and structural variability, the theoretical complexity of the protein universe is essentially limitless. Furthermore, sequence space currently appears to be growing randomly, in that common protein functions (if not folds) can exist even whilst sequence similarity approaches that of random sequences[4,5].

However, despite the complexity and specificity inherent in each protein, the diversity of life that we see around us is not a product of individual proteins, but rather of the interactions between them. Indeed, the function of a protein is usually stated in terms of its connections with other biological molecules. In the case of enzymes, interactors are typically small metabolites, but most since proteins are not enzymes, the majority of biologically interesting interactions are simply with other proteins. Most of these protein-protein interactions are fleeting, arising as a result of dense intracellular crowding, and may or may not be functional. Often though they last longer, producing stable protein complexes, the formation of which is generally essential to the proper function of the constituent protein subunits. This thesis is concerned with the assembly of such protein complexes, and asks how and why they form, and what the wider biological implications of these assembly processes are?

Before we go on, it is helpful to define some terms: with respect to the quaternary structure of a protein, there are three classes that can be used to encompass all possible structures. To start with, a

single protein chain that exists without forming stable interactions with other proteins is known as a monomer. More common is the case where interactions occur between identical copies of a single protein species; in this case the resulting complex is known as a homomer. Finally, heteromers are formed from groups of distinct, non-identical proteins. In addition to these three categories, I will also refer to the stoichiometry of complexes - that is, the numbers and ratios of different subunits that make up the whole.

Stable protein complexes are common, whether homomeric or heteromeric. Given this, an important question is: what evolutionary processes give rise to protein complexes? A major contributor, particularly to homomers, is likely to be genetic drift. This idea was first laid out in an influential paper by Michael Lynch[6]. In this work he describes a simple model in which transitions between multimeric states are represented as a Markov process, with transition probabilities being dependent on mutation rate and selection pressure. The implication arising from this work is that, under neutral to modest selection pressure - which is the case for the large majority of eukaryotic genomes - mutations causing homomerisation of a given protein will arise regardless of the direction of selection. Multiple studies have now been published that strongly support this idea[7–9].

Genetic drift is a major driver of evolution, particularly in multi-cellular eukaryotes such as ourselves, and non-adaptive evolution must therefore be viewed as an important null hypothesis before turning to adaptive explanations[10,11] for the formation of protein complexes. However, there are numerous benefits provided by a modular system of protein formation. For example, when considering the metabolic cost of synthesising proteins, it may be more effective to split a large protein into parts, so that errors in translation are restricted to smaller units. Since the error rates in gene expression are such that they present a major challenge to the viability of life, this reason seems to be particularly plausible[12]. More spectacular examples of modularity in action can be seen with complexes such as ATP synthase, the ribosome, or the cell translation machinery, all of which it is hard to imagine existing in forms other than their present ones.

Another potentially adaptive and widespread phenomena arising from the formation of protein complexes is that of allostery. The original definition of the term given by Monod, Changeux and Jacob[13] referred to modulation of protein activity by small molecules binding away from the active site, but this has since been extended to include cooperative effects between proteins. The classic example of allostery is haemoglobin, in which the binding of oxygen to one subunit increases the binding affinity of neighbouring subunits by propagating structural changes through the subunit binding interfaces[14]. Intriguingly, whilst one might expect beneficial allosteric mechanisms to be uniformly conserved, it turns out that the opposite is often true; as a case in point, haemoglobin is known to differ mechanistically across species[15–17].

1.2 A BRIEF HISTORY OF RESEARCH ON PROTEIN COMPLEXES

The tendency of proteins to form complexes and the functional implications of this behaviour has been recognised since the earliest days of molecular biology. Though it is unclear who was the first to explicitly note their existence, it seems likely that interest in protein complexes arose in tandem with investigations into the nature of viruses. In 1935, W. M. Stanley reported the isolation of ‘a

crystalline material which has the properties of tobacco-mosaic virus' (TMV), and demonstrated that this material was predominantly composed of protein[18]. However, it is not obvious whether or not he understood the implications of finding such a structure for proteins beyond those comprising the TMV capsid. Either way, this period in time marks a turning point for the field of biology, and over the next few decades much of the groundwork was laid for our current understanding of protein structure.

As if to usher in the era, 1944 saw the publication of Erwin Schrödinger's classic book: 'What is Life?'[19], which inspired a number of scientists, particularly physicists, to try their hand at biology. Amongst these were names such as Francis Crick, James Watson and Maurice Wilkins, who were are best known for their discovery in 1953 of the structure of DNA[i]. Also familiar with the book, though not so enamoured with it[20], was Max Perutz, who was at the time working on haemoglobin. By this point, it was clear that many proteins were multimeric assemblies, and by 1955 the TMV capsid had been explicitly described as a self-assembling homomer comprised of several thousand identical subunits[21]. All that remained for the study of proteins and their complexes to begin in earnest was the production of the first structural models. This feat was achieved before the end of the decade by John Kendrew and Max Perutz; first with monomeric myoglobin[22], and shortly thereafter, tetrameric haemoglobin[23]. In solving these near-atomic resolution structures, they opened up the door to the new field of structural biology.

Following the Second World War, technology improved rapidly, and during this period structural biology was one of the most productive fields in all of science; X-ray crystallography in particular deserves special mention, having led to 14 Nobel Prizes since 1914. Of these, uncovering the structure of the ribosome - a huge complex consisting of dozens of protein and rRNA subunits - is perhaps the crowning achievement[24–26].

However, whilst X-ray crystallography was in its heyday, other fields were not silent. A classic molecular biology technique that appeared in the late 80's was the yeast-2-hybrid assay (Y2H)[27], in which two proteins of interest are fused to a DNA binding domain and a transcriptional activator domain, allowing binary interactions (or lack thereof) between the proteins to be detected by the expression of a reporter gene. This assay has been enormously successful, with the original paper having been cited nearly 7000 times since publication; despite its age it is still relevant today, notably through use in a high-throughput manner[28]. However, though simple and cost-effective, there are inherent limitations to the technique: most obviously, the involvement of bulky reporter domains risks disrupting or preventing subtle interactions between many proteins. As a result, approaches using mass spectrometry have largely superseded Y2H as the method of choice for quantitative studies of the interactome.

Mass spectrometry is at least as old as X-ray crystallography, but its use in the study of protein complexes did not become possible until the development of soft matrix-assisted laser desorp-

[i] Infamously, Rosalind Franklin was snubbed by Watson and Crick, who did not properly credit her for her essential work in producing the X-ray diffraction patterns that were used to solve the structure. After her experiments on DNA, Franklin was also involved in pioneering work using crystallographic electron microscopy to investigate the structure of viruses. Sadly, she died in 1958 at the age of 37, before achieving the recognition she deserved, but this work later led to her protegé, Aaron Klug, winning the 1982 Nobel Prize for Chemistry. It seems possible that had she lived, Franklin would have been in line for two Nobels, so maybe she gets the last laugh in the eyes of history.

tion/ionisation (MALDI) and related techniques by Karas, Bachmann, Hillenkamp and Tanaka[29,30]. A short while after these breakthroughs, electrospray ionisation (ESI) also became available for use with proteins[31]. Both MALDI and ESI are now essential tools in biology, and by coupling mass spectrometers with liquid chromatography (LC) and affinity purification it is possible to infer the existence of protein complexes from large scale protein interaction data.

Computational biology was launched in the '60s by the prescient efforts of researchers such as Margaret Dayhoff and Russell Doolittle, and has now developed into a mature field capable of tackling diverse and important problems, including the modelling of protein structure. The first steps in computational biology were predominantly concerned with producing tools and databases for protein sequence analysis (at a time when less than 100 proteins had been sequenced![32]), but very quickly work began on force fields for modelling simple chemical structures[33]. Michael Levitt began trying to apply these early force fields to the problem of protein structure in 1968 during a stay in John Kendrew's lab at the LMB[34]; eventually this work, along with that of others in the community, led to the development of CHARMM[35] (Chemistry at HARvard Molecular Mechanics) and AMBER[36,37] (Assisted Model Building and Energy Refinement), both of which are still amongst the most popular force fields in use today.

However, molecular dynamics is computationally intensive, and it has only recently become useful as far as protein complexes are concerned. In contrast, in the early 21st century, when computers were not yet powerful enough to model large proteins, the number of sequences available to researchers was beginning to explode, and this wealth of data made homology modelling possible[38,39]. This has had a huge impact on our ability to predict structures, and currently aids a variety of structural methods by providing templates to guide model building.

Protein complexes as distinct entities are important, but the proteome as a whole is a highly dynamic ensemble, and our understanding of either is not complete without considering the processes behind assembly and disassembly of complexes. The following chapters are predominantly concerned with assembly, and as we shall see, the process is of great biological importance. Many phenomena can be better understood by taking assembly into account - for example, the evolution of gene order in bacteria (chapter 2), or the attenuation of protein levels observed in aneuploid cells (chapter 4). However, it has only been recently that the technical developments highlighted above have reached a level of sophistication where they can be used to investigate these processes. Since the work in this thesis is built upon these techniques, this chapter provides a broad overview of the current state-of-the-art in structural, non-structural, and computational methods for investigating protein complexes.

1.3 STRUCTURAL CHARACTERISATION OF PROTEIN COMPLEXES

Being able to visualise something when we are studying it is an invaluable aid to our understanding, and this is certainly true for proteins. In this regard, the field of structural biology is a satisfying one, as it enables us to picture (however unrealistically) molecules that exist at a scale far below anything in our day-to-day experience. Though the field was given life by X-ray crystallography, cryo-EM and NMR spectroscopy are nowadays equally important, with each technology occupying its own niche

in terms of the type of problem it is best-suited to. Together, these technologies are responsible for the solution of many thousands of protein structures, and have together have revolutionised biology, medicine and the pharmaceutical industry.

1.3.1 X-ray crystallography

X-ray crystallography was the first method to make the field of structural biology a reality, bringing together three separate technologies, each important in its own right. These technologies include: methods for overexpression and purification of proteins, the production of powerful X-ray sources, and computational methods for solving X-ray diffraction patterns. By and large, the ways in which X-ray crystallography can be used to determine protein structure are the same for monomeric proteins and those which form complexes. There are however some important differences and additional difficulties that need considering in the latter case. Although cryo-EM seems poised to overtake X-ray crystallography as the method of choice for the solution of large heteromeric structures, there have been a number of exciting developments in crystallography that look set to ensure its future for many years to come. In the following section I will highlight of some of these advances, and attempt to give a summary of the current state of the field.

Protein expression, purification, and crystallisation

Acquiring samples of purified protein is a requisite first step for almost all of the methods discussed in this chapter, and X-ray crystallography is no exception. Though I will describe the basic principles with the production of crystals in mind, much of what follows is very general and is applicable to many other techniques; for further reading, general reviews on the topic of protein purification can be found in the bibliography[40–42].

A typical procedure for the expression of protein for crystallisation involves transforming *E. coli* with a plasmid containing your protein of interest, usually under the control of a strong, inducible promoter[43]. For monomeric bacterial proteins this system is simple and easy to use, but expressing heteromeric protein complexes is often considerably more challenging, particularly those of eukaryotes. The key difficulty in the expression of heteromers lies in the production of sufficient quantities of pure sample, as in non-native host systems protein complex assembly is often inefficient or simply incomplete, making purification and subsequent crystallisation challenging. For eukaryotic proteins, this is compounded further by the fact that most undergo alternative splicing and other post-transcriptional or -translational events, the machinery for which is generally lacking in bacteria.

Prior to any bench work, improvements in the cellular yield of bacterial heteromers can be achieved by carefully considering the design of the expression vector in light of the assembly pathway of the protein complex in question. As will be discussed in detail in chapter 3, the order of genes within operons is under selection to match the assembly order of protein complexes[1]. It has been demonstrated experimentally that taking this fact into account can markedly increase complex assembly efficiency, and that yields of heteromers in their fully-assembled native state can be improved by using the native operon structure in expression vectors[44,45].

When purifying protein complexes there is a tradeoff between obtaining highly pure samples and ensuring that the intermolecular bonds between subunits are not disrupted. Though the diversity of methods for protein purification is bewilderingly high, in practice most methods suitable for protein complexes are variations on affinity purification. Here too, careful experimental design can pay dividends, and when possible it is generally preferable to produce bait proteins that are expressed at endogenous levels[ii]. Ideally, the number of purification steps would be limited in order to retain as much protein in its native state as possible, but in practice multiple purification steps are often required before the sample is pure enough to crystallise. Methods such as dynamic light scattering[47] can be used to assess sample purity and readiness for crystallisation.

In most cases it will be necessary to tailor the expression and purification process to the protein complex of interest. Depending on the orientation of subunits within the structure for example, different subunits may make better or worse bait proteins, as will N- or C-terminal purification tags. Similarly, some complexes may be disrupted by the presence of metal ions, in which case other beads, e.g. those coated in calmodulin, may be more suitable. Ultimately, though there has been progress towards high-throughput expression and purification pipelines[48], much of this work still relies on trial and error informed by the expertise of individual structural biologists and research technicians.

Surprisingly however, the main bottleneck in X-ray crystallography is crystallisation, despite having been largely automated by the development of screening robots. Having said that, there have been some important methodological developments in the crystallisation of membrane proteins, which will also be useful for many membrane complexes. For example, an exciting new method - X-ray solvent contrast modulation - has recently been used to visualise the interaction between membrane proteins and the phospholipid bilayer[49]. However, this method does not do away with the requirement for good quality crystals, and these are still largely obtained through trial and error - beyond a few general rules of thumb we still do not have a good understanding of how different proteins will behave under different crystallisation conditions.

Diffraction pattern acquisition

Once suitable crystals have been obtained however, the main hurdle has been hurdled and image acquisition can begin. In contrast to earlier steps, enormous progress has been made in this domain since William L. Bragg first demonstrated X-ray diffraction from sodium chloride crystals in 1913[50]. By far the most important development in the field has been that of synchrotron X-ray sources. Synchrotrons are able to produce X-rays at far higher intensities than traditional sources, and as such greatly reduce the time it takes to produce diffraction patterns. Technical properties of the beamline can also be manipulated, for example narrowing it in order to focus on the best quality region of the crystal, thus improving the quality of the resulting diffraction pattern.

[ii] Somewhat counterintuitively, increasing the abundance of a single subunit may actually decrease the yield of the native complex. To understand this, imagine a trimer, assembled linearly as follows: A-B-C. If the concentration of subunit B were to be doubled, the resulting imbalance in stoichiometry would lead to A and C being preferentially sequestered in the form of A-B and B-C dimers, which are incompatible with the original trimeric structure. The idea that differentially modulating subunit expression within complexes can be deleterious is known as the balance hypothesis[46], and will be discussed in more detail in coming chapters.

More recently, X-ray free electron lasers (XFELs, figure 1.1) have begun to make an appearance in structural biology. It is hard to overstate the impact that this technology will have on the field, since XFELs are capable of producing peak beam energies approximately ten orders of magnitude greater than current 3rd generation synchrotrons[51], and in doing so enable a radically different approach to crystallography. The principle benefit of all this additional power is that the time needed to generate a diffraction pattern is drastically reduced: from hours to femtoseconds. A crystal in the path of such high-energy photons will be vaporised almost instantaneously, but since the diffraction pattern will be obtained faster than the sample is destroyed, this proves not to be a problem - a fact first noted by Neutze et al.[52], giving rise to the term 'diffraction before destruction'. Of course, this generates a need for a great many crystals in order to obtain diffraction patterns of the structure from all angles, but this too is not a major issue, since these crystals only have to be a few nanometres in size. In fact, since nanoscale crystals are far easier to grow, the method also circumvents the tedious trial and error process of producing the larger crystals needed for use with traditional X-ray sources.

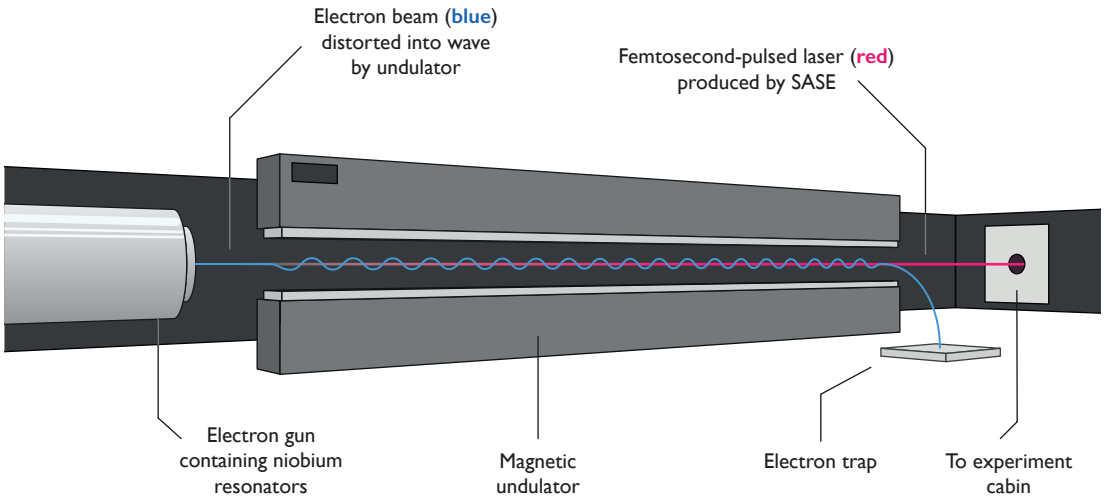


Figure 1.1.: X-ray free electron lasers
An XFEL produces high energy X-rays by a process known as self-amplified spontaneous emission (SASE). An electron bunch is accelerated close to the speed of light using superconducting niobium resonators. When this passes through the undulator, the wiggling motion induced by the magnets causes the electrons to emit photons. As these photons are travelling only slightly faster than the electrons, they interact with the electrons as they catch them up at each period in the undulator. Over the length of the undulator, this causes the electrons to bunch into very thin disks, which emit intense, synchronised flashes of X-ray laser light. These femtosecond X-ray pulses are then guided into the experiment cabin, where they encounter a stream of protein nano-crystals, producing diffraction patterns from each one.

Structure determination

Interpretation of the crystal diffraction pattern required the solution of a long-standing challenge in the early days of X-ray crystallography, namely the phase problem[53]. The phase problem exists due to the fact that, whilst diffraction patterns capture the amplitude of diffracted photons from a

crystal (seen as the intensity of spots on the photograph), the phase of those photons is lost in the process of image acquisition. Unfortunately, it is the phases of the diffracted photons, rather than their amplitudes, that carry most of the information about the underlying crystal structure. Indeed, it was the eventual solution of this problem by Max Perutz that was the key to his and Kendrew's determination of the first protein structures.

Perutz's breakthrough came when he realised that a technique previously used for phasing crystals of much smaller molecules could also be applied to proteins. This method, known as isomorphous replacement (IR)[54], involves soaking the crystal in a solution containing heavy metals. Crucially, the incorporation of heavy metals into the crystal does not significantly alter its structure, and as a result, the position of spots in the diffraction pattern remain almost unchanged, whilst subtle differences in their intensity point to the location of the heavy atoms. This provides an essential reference point for calculation of the missing X-ray phases.

For large protein complexes, polynuclear metal clusters are often used in place of individual heavy atoms because of their particularly electron density and associated isomorphous or anomalous scattering signal[55]. This approach has recently been used to good effect in solving the structure of the notoriously difficult mediator complex[56]. However, different methods for solving the phase problem have been established in addition to IR, most notably multiple wavelength anomalous diffraction[57] (MAD). This method operates on different principles to IR but is popular since it is limited only by the quality of the diffraction pattern provided to it.

As a consequence of the ever-expanding number of structures in the Protein Data Bank[58] (PDB) and the widespread availability of sequence data, it is often possible nowadays to avoid de novo phasing altogether. Molecular replacement by homology modelling (discussed later) makes use of the fact that closely related sequences generally have similar folds, and therefore can be used as a template to guide brute-force calculation of diffraction pattern phases. There are currently several programs that automate this process - for example, Phaser[59], which is available within the widely used CCP4 software suite[60].

1.3.2 Cryo-electron microscopy

X-ray crystallography has been, and will continue to be, an enormously useful too for investigating proteins and protein complexes. However, a recent resurgence in cryo-EM has had a transformative effect on structural biology - particularly on our ability to solve the structures of large protein complexes above 300 kDa in size. Its unique affinity for large complexes is especially convenient since these are often prohibitively difficult to crystallise, in large part due to compositional heterogeneity of the purified samples, which cryo-EM can more easily handle. The two methods are therefore highly complementary, and indeed many structures are solved to good resolution by a combination of the two - cryo-EM for the coarse-grained structure, and X-ray crystallography for atomic resolution detail of individual subunits. Likewise, NMR also has difficulty handling large complexes, and thus can be used effectively in combination with cryo-EM.

As interest in cryo-EM increases (in March 2017 the Wellcome Trust announced a £20M grant for cryo-EM equipment in several UK laboratories), there are signs that single-particle cryo-EM

is making incursions into the size and resolution niche currently occupied by X-ray crystallography. Illustrating this, two important symbolic barriers were recently broken in a 2016 Cell paper describing the structures of two homomeric complexes: isocitrate dehydrogenase and glutamate dehydrogenase[61]. The former weighs in at just 93 kDa, and is the first single-particle cryo-EM structure of a <100 kDa complex, whilst the latter was resolved to 1.8Å, breaking the <2Å resolution barrier. As we shall see, the remarkable technological achievements displayed in this paper and several others have been driven by dramatic improvements in the two key areas of image acquisition and image processing[62].

Image acquisition in single-particle cryo-EM

The first major development behind cryo-EM's current flourishing came with the replacement of photographic film by digital direct electron detectors, specifically monolithic active pixel sensors (MAPS). It was not until relatively recently that digital detectors came into widespread use, and until the arrival of MAPS, film was the medium that achieved the best possible detective quantum efficiencies (DQE)[63]. DQE is a measure of the signal to noise ratio that can be achieved relative to an ideal detector[64], and is defined as follows:

$$DQE = (S/N_{in})^2/(S/N_{out})^2$$

Where S/N_{in} and S/N_{out} are the input and output signal-to-noise ratios respectively; a DQE of 1 would imply that the detector was not responsible for any noise in the image. For reference, film has a DQE of around 0.3, whereas the current state-of-the-art digital detectors achieve roughly twice that.

Ultimately, DQE is the most important factor in choosing whether to use film or digital detectors, but now that MAPS have surpassed film in that regard, several other compelling advantages of digital detectors can exploited. From a practical standpoint, they are significantly faster to use than film, since images can be viewed immediately after collection and their acquisition can be automated. They can be used to produce high frame-rate videos, enabling them to be run in counting mode, where instead of integrating the signal produced by each incident electron across all the pixels in which a charge was registered, only the pixel with the highest charge is counted[65]. This is conceptually similar to the way in which certain microscopy techniques achieve super-resolution images, and the company Gatan has recently brought this idea to market a with a dedicated super-resolution mode for their K2 Summit detector.

One exciting new technology which is beginning to make its presence felt is the Volta phase-plate, which can be used to directly modulate phase contrast during image acquisition. In order to be able to correctly distinguish between different particles in the sample it is important to have good contrast in the images. Unfortunately, the method by which this contrast is currently changed relies on defocusing the image slightly and as a result, if greater contrast is required, it comes at the expense of resolution. The Volta phase-plate circumvents this issue by modulating the phase directly, without affecting the focus of the image[66]. Though the principle has been around for some time, it was not until recently that various practical issues were solved, enabling Bai and colleagues

to produce a 3Å structure of the 20S Proteasome, thus matching the resolution achieved by defocus methods[67]. Most impressively, the same group has just this year published a 3.2Å structure of the 64kDa haemoglobin molecule[68].

Image processing and structure determination

A second important factor in cryo-EM's success has been the appearance of better image processing software, which has enabled researchers to get the most out of the concurrent improvements in imaging hardware. In addition to improving resolution, the emergence of electron detectors capable of producing high frame-rate videos in counting mode has a secondary benefit, in that it enables beam-induced motion blurring in the images to be corrected computationally, a feat first achieved by two groups almost simultaneously in 2013[69,70]. Since the reduction in signal quality incurred by beam-induced movement is around five-fold if uncorrected[71], this was a major breakthrough at the time, and is now standard protocol that can be carried out with the popular image-processing program RELION[72,73].

Computational progress has also been essential for image classification (figure 1.2). In single-particle cryo-EM, individual protein complexes are fixed in random positions and orientations in the flash-frozen sample - to determine the structure, each particle captured in the imaging process must first be categorised according to its orientation. For symmetrical structures, the number of particles required in the image is usually considerably lower than for asymmetric structures, since multiple axes of symmetry effectively make many of the observed orientations redundant. This has the effect of increasing the effective number of images of symmetric particles, and conversely, ensures that the solution of asymmetric structures is more challenging.

Dealing with structural and compositional heterogeneity is a related problem, which arises from imperfect sample purification or the presence of different functional states of the complex. Computational approaches for dealing with this arrived in 1998 with a maximum likelihood method for classifying two-dimensional (2D) images[75]; 3D classification methods, being much more computationally intensive, did not appear till later, but are now an area of active development, since they are currently one of the major bottlenecks in structure determination[76–78]. In practice, multiple rounds of image classification and refinement are usually carried out, beginning with removal of low quality particles, followed by 2D and 3D image classification and finishing with polishing steps.

Cryo-electron tomography

Although single-particle cryo-EM offers good resolution without the need for crystallisation, it still requires that the protein of interest be purified first, thus ruling out many of the protein complexes present in the cell, including those embedded in the cell membrane. Cryo-electron tomography offers an attractive alternative in these cases, as it allows imaging of protein complexes in their native environment, albeit with a significant reduction in the resolution achievable. By and large, the processes involved in cryo-ET are similar to those of cryo-EM, with the key difference being that one acquires images by rotating the sample through a range of different tilts, rather than relying on

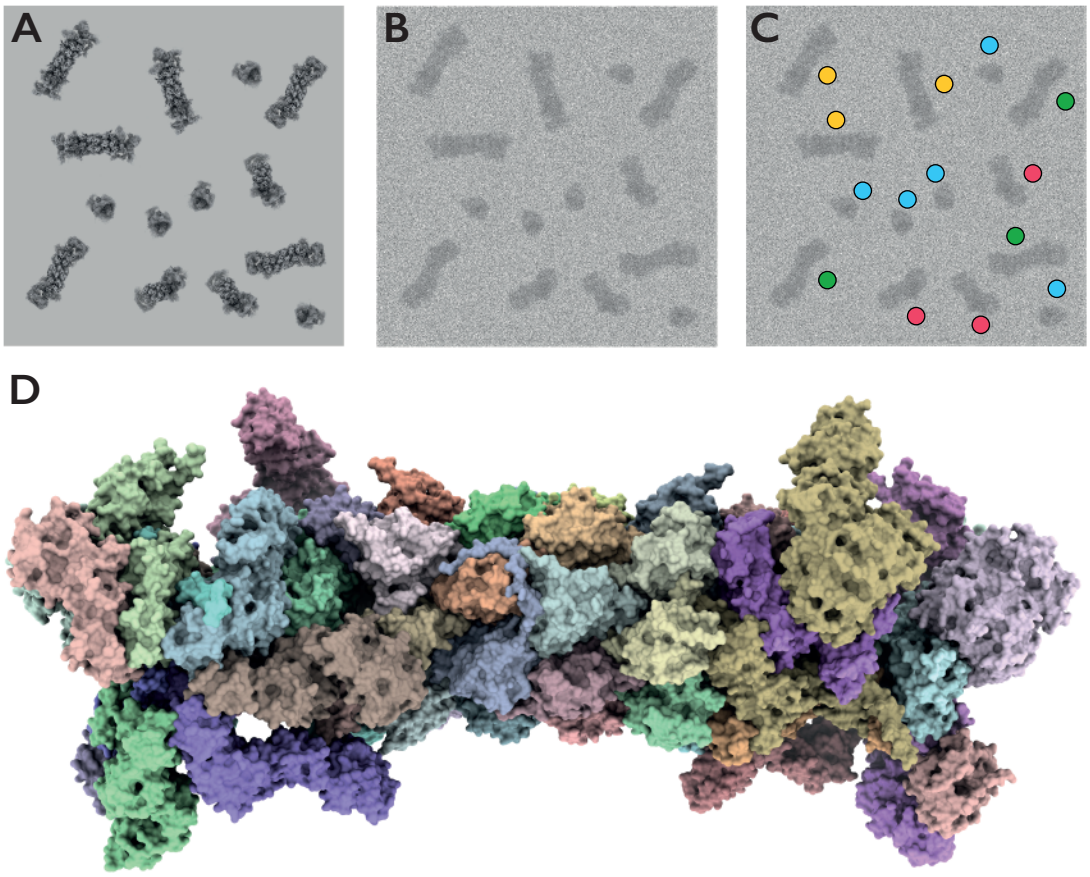


Figure 1.2.: Image classification in single-particle cryo-EM.
(A) Theoretical electron micrograph of the human 26S proteasome produced by a detector with DQE = 1.0. (B) Image produced by detector with lower DQE, resulting in noise and phase contrast issues. (C) 2D image classification of proteasome particles into categories corresponding to their orientation. (D) Fitted 3.8Å resolution model. Produced using K2 summit detector and processed in RELION; PDB ID: 5T0C, EMDB ID: 8332[74]

the protein naturally being present in many different orientations. As an aside, this tilting method is also used to produce images in electron crystallography[79].

By reconstructing the set of images produced from these different tilts, a tomogram of the structure of interest can be built. Because the exact orientation of the sample is known for each image, confounding factors such as other proteins or biological structures can be removed from the image, which would not be possible if one were to attempt single-particle cryo-EM on a non-purified sample. The downside to this approach is that the sample can only be tilted up to a certain point, as the effective thickness of the sample in the path of the electron beam increases with the angle of the sample. As a result, there is always a 'wedge' of data missing from the set of images of a complex, seriously limiting the resolution achievable from a single structure.

However, an important development of cryo-ET is subtomogram averaging, otherwise known as single-particle tomography[80]. Here, multiple tomograms of different particles in the sample are produced, and then averaged in similar fashion as for images in cryo-EM[81]. This averaging process can fill in the missing wedges in the data, provided the proteins in the sample are present in a sufficient variety of orientations[82]. Though the technique is not yet able to reliably achieve atomic resolutions, it is not far off[83], and the lure of imaging protein complexes in their natural environment will almost certainly ensure its continued development.

1.3.3 Nuclear magnetic resonance spectroscopy

Many biologically important protein complexes exist in a dynamic ensemble of conformational states, or contain subunits that only interact very weakly with each other. Such complexes do not lend themselves well to characterisation by crystallographic or cryo-EM methods, which can only resolve a single structural conformation at a time. Nuclear magnetic resonance spectroscopy is well suited to investigating these cases as the proteins are visualised in solution, rather than crystallised or frozen. On the other hand, NMR has traditionally struggled to resolve structures beyond 30kDa due to the fact that the relaxation of nuclear spin orientations is very efficient for large, slowly tumbling molecules. This has the effect of broadening the peaks observed in NMR spectra and, coupled with the fact that large molecules naturally produce more complex spectra than smaller ones, ensures that using NMR to study protein complexes is challenging.

Solution NMR spectroscopy of multi-subunit protein complexes

An essential tool for investigating large complexes is transverse relaxation-optimized spectroscopy[84] (TROSY), which uses constructive interference between different relaxation effects to improve the resolution of chemical shifts. Equally important is the use of deuterium (^{2}H) labelling[85]. Like TROSY, this improves resolution by increasing the relaxation time of molecules. A further extension of these concepts is methyl-TROSY, which makes use of isotopically labelled $^{13}C^{1}H_3$ methyl groups set against a highly deuterated background. Because methyl groups produce especially intense resonances, they are easily identifiable within NMR spectra, and furthermore they are well dispersed within nearly all protein structures[86]. Using this technique it is possible to resolve proteins with molecular weights into the low hundreds of kDa, overlapping slightly with the lower

limits achievable by cryo-EM.

For yet larger protein complexes, or those with more heterogeneous structures, the complexity of the spectra itself becomes the limiting factor, rather than the spin relaxation rates. In these cases, clever use of isotope labelling can often simplify matters considerably (for a nice review, see Zhang and van Ingen[87]). Segmental labelling is one such example, in which isotopically labelled regions of the protein are spliced in using inteins or sortases[88]. Unsurprisingly, this is fraught with technical difficulties, but despite these the method has been used to great effect in studying large protein structures, from the 0.6MDa ClpB disaggregase chaperone[89] to prion protein amyloid fibrils[90].

Solid-state NMR

The latter of these studies - characterising amyloid fibrils - made use of solid-state NMR spectroscopy. As the name suggests, this requires sample in a solid state, which is then spun rapidly inside the magnetic field, as opposed to the molecule of interest being free to tumble in solution. This is possible because of a quirk of NMR that leads to the delightfully named 'magic angle spinning' technique[91,92]. When the sample is tilted at the magic angle θ_m relative to the external magnetic field (such that $\cos^2\theta_m = \frac{1}{3}$), the peaks on the NMR spectra become much sharper, enabling structure to be determined. Magic angle spinning in effect mimics the natural tumbling of molecules in solution, but since the rate of 'tumbling' is no longer dictated by the size of the macromolecule being observed, solid-state NMR can be used to probe much larger structures (e.g. amyloid fibrils).

It is also well suited to studying membrane-embedded protein complexes, as a result of the fact that proteins in lipid bilayers are by default oriented in a single direction. Through careful sample preparation, this natural orientation can be preserved during the course of the NMR experiment, allowing high-resolution spectra to be produced directly from the sample by aligning it at the correct angle to the external magnetic field[93]. A number of impressive complexes have recently been solved using both oriented-sample methods and magic angle spinning[94–96].

1.3.4 Electron paramagnetic resonance spectroscopy

Electron paramagnetic resonance (EPR) spectroscopy is related to NMR spectroscopy, but differs in that it detects species with unpaired electrons, rather than nuclei more generally. This is possible due to the Zeeman effect[97]: In a static magnetic field B_0, the magnetic moment of the unpaired electron aligns with the field in either a parallel or anti-parallel direction, with an associated energy level for each direction. An electron moving between these two states will emit a photon with energy corresponding to the difference between the two levels, and in this way the species containing the electron can be detected.

The precise energy of the emitted photon is dependent on several factors, including the strength of B_0 and, more importantly, the local electronic environment of the species in question. In particular, hyperfine coupling leads to characteristic splitting of peaks in the spectra, allowing functional groups such as methyls to be easily identified.

The difficulty in applying EPR spectroscopy to the study of protein structure lies in the relative

scarcity of unpaired electrons in biological molecules. Proteins with naturally occurring paramagentic species such as the haem iron in haemoglobin make good candidates, or alternatively the problem can be overcome by site-directed spin labelling[98,99]. These labels can be tailored to the problem in question; for example measuring responses to pH changes, or behaviour under reducing conditions. For a review of the topic see Klare, 2013[100].

Combined with site-directed spin labelling, EPR is a powerful tool for investigating complex macromolecular systems and process. One such example is to demonstrate that the partially disordered mitochondrial ATPase inhibitor underergoes pH-dependent conformational changes in its dimeric state[101]. Perhaps more impressively, the PsaC subunit in photosystem I has been monitored through a number of assembly steps, showing several conformational changes corresponding to the binding to the complex of both PsaC and other subunits[102].

1.4 NON-STRUCTURAL CHARACTERISATION OF PROTEIN COMPLEXES

Structural methods are an essential tool for describing and understanding protein complexes, but by definition they have fairly limited applicability beyond providing structural information. If we wish to have a complete characterisation of a given complex, structural methods alone are insufficient, as there is a great deal of useful information that cannot be determined solely from snapshots of a given conformational state. For example, the pathways by which protein complexes assemble, or the degree to which their subunit composition varies under different conditions.

It is important to try and understand the behaviour of protein complexes within the context of the wider proteome. Many cellular phenomena can be best explained as emergent properties of the complete network of protein-protein interactions, and mass spectrometric techniques have been particularly useful in quantifying the protein interactions that take place within cells. Through technological innovation and clever experimental design, mass-spectrometry has proven to be highly versatile, and has been used for a number of different purposes, including elucidation of protein complex assembly pathways[103,104], investigations into the evolutionary history of complexes[105], and generation of richly detailed interactome datasets[106].

1.4.1 Native mass spectrometry

The arrival of soft ionisation MS techniques in the '80s was of critical importance for the study of protein complexes, as it allowed delicate non-covalent interactions between proteins to be preserved in the gas phase, making it possible to study intact protein complexes via MS. Combined with the later development of time-of-flight mass analysers, this method became known as native MS. Because native MS does not interfere with the intermolecular bonds between protein complex subunits, it can be used to study properties such as stoichiometry, compositional heterogeneity, and dynamic processes such as assembly or disassembly.

Electrospray ionisation mass spectrometry

The ionisation method of choice for native MS is currently ESI, as MALDI requires the sample of interest to be mixed with a matrix, which is then ionised using lasers. This matrix is usually formed

from crystallised organic acids, and as such is generally too harsh for complexes to be maintained in their native state, with a few exceptions[107]. In contrast, ESI uses the sample as is, and ionises it by passing it through a narrow glass capillary to which a high voltage is applied, causing the charged sample to be aerosolised as it leaves the capillary. Through successive Coulomb fission events and evaporation of solvent from the sample, the ions in this mist rapidly enter the gas phase as they move towards the mass analyser.

Another important benefit of ESI over MALDI is that it produces multiply charged protein ions with regularity[108]. This is useful when coupling ESI to tandem MS, where the protein sample is first analysed in its native state, before being fragmented and subject to a second round of mass analysis. Single charge proteins produce little useful information upon fragmentation as only a single peptide fragment will be charged, essentially wasting much of the protein. Having multiple charges per ion also reduces the corresponding m/z ratio. This is important when investigating larger proteins and protein complexes, since (historically at least) the operative range of quadrupole mass analysers has been limited to about 4000 m/z. For this reason, time-of-flight mass analysers have been the mainstay of native MS for many years (reviewed by Radionova et al.[109]), since they have good resolving power and sensitivity over a much wider range than traditional quadrupole analysers. In 2005 however, Orbitrap analysers became available[110], and subsequent developments since then have pushed the limits of their operative mass range into the tens of thousands m/z.

Another hugely important development in ESI came with the introduction of much narrower capillaries in the electrospray devices, leading to nano-ESI[111]. Coupled with lower sample flow rates, this improves ionisation efficiency substantially[112]; equally importantly, it greatly reduces the amount of sample required for each experiment. This enables analysis of proteins that are hard to purify in large quantities, or makes it possible to run experiments investigating dynamic processes that take place over the course of seconds to minutes.

Applications of native mass spectrometry

Due to its low sample requirements and sensitive treatment of the intermolecular interactions, native MS is very versatile. A common and technically straightforward use of the method is simply to determine the constituent parts of a particular protein complex, which can be done via tandem MS[113,114]. The weights of individual subunits from the complex are determined in the first round of mass analyses, with identities being inferred from fragmented peptides in the second round. From this starting point, it is then possible to generate interaction maps based on the weights of peaks corresponding to different subunit combinations, as well as getting an idea of relative binding strengths.

More interesting is the use of native MS in time-resolved studies, for example: following subunit exchange processes between heat-shock proteins[115], observing conformational changes of membrane complexes upon ligand binding[116], and determining protein complex assembly and disassembly pathways[103]. This last example, which is of particular importance to the rest of this thesis, can be achieved by adding different chaotropic agents to the solution containing the intact protein complex, then observing the intermediates that are produced across different concentrations.

1.4.2 Cross-linking mass spectrometry

Cross-linking mass spectrometry (XL-MS) uses chemical cross-linkers to provide distance constraints between different residues in a protein complex. These can either be intramolecular or intermolecular, and as such XL-MS be used to produce low-resolution structural information, particularly on the interfaces between different subunits. It is particularly effective when used in combination with more established structural techniques or together with computational modelling, and as such has become an central part of the new, integrative approach to structural biology[117–119].

Chemical cross-linkers

The power of XL-MS comes from the availability of a wide variety of different cross-linkers that impose specific distance constraints on the interactions being probed. These can be tailored to the question at hand, with the selection of cross-linker lengths placing different constraints on the interactions that can be studied. Similarly, the biochemical specificity of these linkers can be used to look at interactions between specific functional groups. Most commonly used are homobifunctional cross-linkers that join primary amines[120], i.e. lysine residues or N-termini, with spacer arm lengths ranging from ~3Å to ~35Å.

Heterobifunctional linkers (in contrast to homobifunctional ones) allow different groups to be targeted, for example joining amine to carboxyl (aspartate, glutamate, C-termini) groups. More nuanced experiments can be performed by using some of the more exotic linkers that are currently being produced. Heterobifunctional photoreactive cross-linkers such as aryl azides are attractive for in vivo applications since they are inert until photoactivation, at which point they rapidly form non-specific cross-links with chemical moieties in their immediate environment. Photoreactive analogues of some amino acids have also been discovered, enabling incorporation of linkers into the protein sequences themselves[121].

Notable applications of XL-MS

XL-MS is well suited to looking at flexible complexes that cannot be observed using cryo-EM or X-ray crystallography. A good example of this is the family of SMC-kleisin complexes that will be discussed in chapter 5. These complexes are essential for accurate cell division and are formed of heterodimeric, coiled-coil SMC subunits, joined by a disordered kleisin subunit to form a trimeric ring structure that entraps DNA. Several crystal structures of the various the subunit interfaces (minus flexible regions) are available, but thus far XL-MS has been the only method that has had success modelling the topology of the entire complex[122]. Interestingly, cross-links between the two SMC arms suggest that when not encircling DNA the SMC arms are collapsed in on themselves.

A more formidable test of XL-MS comes from the ongoing effort to understand the structure of the nuclear pore complex (see Beck and Hurt, 2016[123]). Due to their enormous size (~120 MDa in humans, compared to ~3.5 MDa for the ribosome) and high degree of compositional variation between species, it is difficult to distinguish between subunits, many of which are paralogues of each other. In such cases, XL-MS can provide essential information about the specific identity of different subunits and their contacts, allowing the identification of ambiguous subunits within

larger cryo-EM electron density maps[124,125].

1.4.3 Affinity-purification mass spectrometry

In its simplest guise, AP-MS enables the identification and quantification of the interaction partners of a given protein. The general principle is as follows: a column containing beads capable of capturing your bait protein is prepared. Native cell extract (though sometimes over-expression of the protein of interest is required) is then washed over the column, leading to the capture both the bait and proteins bound to it via co-immunoprecipitation. The eluent is generally subjected to peptide fractionation, and mass-spectrometry is then used to quantify either the relative or absolute abundances of members of the purified complex. For high-throughput studies, multiple proteins are used as baits, enabling large interaction maps to be generated.

Though conceptually simple, AP-MS is an enormously powerful technique, and one that deserves more attention than it is going to get here. Fortunately, several reviews have been written on the topic, and I therefore direct the reader to these[126–128]; in particular, that by Morris and colleagues is excellent[127]. For the sake of brevity, the paragraphs that follow are limited to just the most important variations of a method that has been instrumental in achieving our current understanding of the protein interactome[105,106,129,130].

Single-step versus tandem affinity purification

There are two approaches to affinity purification in widespread use - single-step and tandem affinity purification[131] (TAP, figure 1.3). In the former, the bait protein is either expressed under completely endogenous conditions and captured using antibodies, or expressed with a tag such as green fluorescent protein[132] and captured using methods appropriate to the tagging system. In contrast, TAP makes use of a unique TAP-tag, which consists of a protein A domain and a calmodulin binding peptide, linked by a Tobacco etch virus (TEV) protease cleavage site. This tag enables a two-step purification procedure that results in stringent purification of complexes, though sometimes at the expense of weak but specific interactions.

TAP necessarily requires tagging of the bait protein, but in the single-step procedure it is possible to avoid this if desired, in which case it is referred to as endogenous purification. The are some straightforward trade-offs to consider when deciding whether to use endogenous or tagged proteins: For non-tagged baits, you have the benefit of capturing the protein in its native state. However, this comes at the substantial cost (both in time and money) of having to raise specific antibodies against the protein in question. Furthermore, there are difficult issues associated with cross-reactivity and specificity when using antibodies, particularly in studies where multiple proteins are being targeted. Though there are methods that attempt to deal with these issues (most notably QUICK[133]), in the majority of large-scale studies prior tagging of bait proteins is likely to be more practical.

When using AP-MS to carry out interactome studies there are some fairly compelling advantages to using single-step procedures over TAP. The purpose of TAP is to remove as many contaminants or non-specific interactors as possible from the purified protein complex. This was essential in the early days of mass spectrometry, since it was not possible to quantify protein abundances, and thus

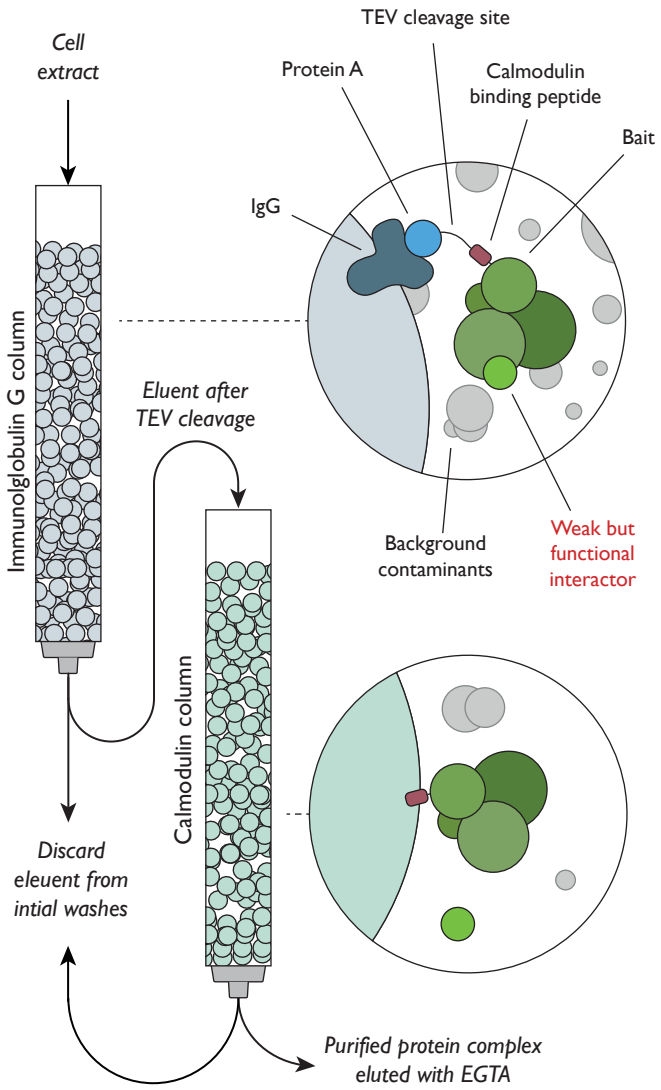


Figure 1.3. Tandem affinity purification protocol
TAP differs from single-step purification procedures in that it requires two distinct washing steps. This is possible due to the TAP tag, which consists of a protein A domain linked to bait protein via a TEV cleavage site and a calmodulin binding peptide. Cell extract is passed through the first IgG column, which captures the bait protein via the protein A domain. By adding TEV protease to the column, the bait protein and it's interactors are released from the column. This eluent is then added to a second column containing calmodulin beads to which the calmodulin binding peptide attaches. Addition of ethylene glycol tetraacetic acid (EGTA) causes the protein complex to be released from the beads. Due to the fairly intense washing process that the protein complex undergoes, there is a fairly high chance of weak interactors being removed in addition to non-specific contaminants. Single-step procedures are more likely to retain these interactions at the expense of overall sample purity.

contaminants in the sample would be erroneously annotated as members of the protein complex. However, there have been enormous improvements in the sensitivity of mass spectrometers since TAP was first described; with these improvements, particularly in the area of label-free quantification (LFQ), it has become possible to discriminate between proteins present at biologically significant concentrations and background noise. In doing so, the importance of weak, non-obligate interactions between proteins has become more apparent[106,134]. Since TAP removes these weak interactors, its utility is becoming increasingly restricted to situations where extremely pure protein is required, e.g. for crystallisation. Therefore single-step procedures combined with accurate quantification should be generally be considered preferable to TAP for large-scale studies. Following this reasoning, a promising new technique named affinity-enrichment purification has recently been described that deliberately uses only very mild washing steps[135].

Quantification of protein abundances

The emergence of quantitative mass spectrometry, via both label-based and label-free methods, has had a transformative effect on the field of proteomics. For our purposes, the principal benefit arising from the ability to quantify protein abundances is that it allows the stoichiometry of protein complexes to be determined. This is essential for distinguishing obligate interactions from transient ones, and more generally for providing a complete characterisation of the complex. The difficulty in using MS as a quantitative tool is that, whilst the location of peaks on the mass spectrum allows identification of peptides, peak intensity alone is not sufficient to determine peptide abundance. Label-based methods such as SILAC[136] (stable isotope labelling and culturing) and iTRAQ[137] (isobaric tag for relative and absolute quantitation) and others[138,139] allow for either relative or absolute quantification.

A significant drawback to label-based methods is their cost, which can be prohibitive. An alternative approach is LFQ, methods for which are based on either spectral counting[140,141] or peak intensity-based algorithms. Spectral counting is a conceptually simple, semi-quantitative approach which has been widely used (and possibly abused[142]). Intensity-based algorithms undoubtedly offer more accurate quantification; for the interested reader, comparative analyses and reviews of several available methods are available[143,144]. One recently developed algorithm of note that has been enthusiastically received by the community is MaxLFQ[145], which is available as part of the larger MaxQuant software package[146].

Inferring protein complexes from interaction data

The data returned from high-throughput techniques such as those we have been discussing are typically presented as lists of pairwise interactions and measures of interaction strength, using proxies such as co-fractionation likelihood in place of true binding affinities. From this starting point, the challenge is then to infer the identity and presence of protein complexes, known or otherwise. Since these pairwise interaction data can be represented in graph form, a common approach to identifying complexes is graph clustering. As might be expected given the generalisable nature of the problem, the literature on this topic is substantial - both specific to protein-protein interactions and otherwise. A comprehensive review of the topic covering both theory and application can be found here[147].

One particularly successful example in this category is the Markov Cluster algorithm[148] (MCL), which is based on the principle that if you take a random walk through a graph, the strongest clusters will be those groups of nodes that you tend to stay in for longest. Despite its age, MCL is still competitive[149], and has been used in a number of excellent interactome papers since its release[105,150]. Other notable mentions include COACH and ClusterOne[151,152], the latter of which is one of the few alternatives to compare favourably with MCL. A very different approach that seems promising is Nano Random Forests, which uses machine learning to identify groups of strongly covarying proteins[153]. As a supervised learning method, it is well suited to investigating known complexes under different experimental conditions, in contrast to those previously mentioned, which are better placed for discovery of novel complexes.

1.4.4 Super-resolution microscopy

Super-resolution microscopy, so named because it breaks the diffraction limit for light microscopy described by Ernst Abbe[154], revolutionised cell biology when it arrived on the scene, enabling visualisation of sub-200nm structures (e.g. virus particles, microtubules). The resolution achievable has continued to drop since stimulated emission depletion[155] (STED) and stochastic emission techniques[156–158] became available, and it is now possible to resolve objects well into the nanometre range, thus enabling the study of protein complexes in their natural environment.

Single molecule localisation microscopy

Though there are now several different ways to beat the diffraction limit, those that are most often used in biology fall under the banner of 'single molecule localisation microscopy'[iii] (SMLM). Common to all of these techniques is the use of photoswitchable fluorescent dyes, which are turned on and off in such a way that only a very small number of molecules are turned on at any given time. This allows for a Gaussian point spread function to be fitted to each fluorescence event, enabling the precise localisation of the molecule responsible. Within the larger SMLM category, there are methods that achieve localisation by shaping the light source used to excite molecules, and those that stochastically switch on and off a sparse, well-separated subset of molecules within the field of view. A widely used example of the former is STED microscopy, which typically achieves resolutions in the region of 50nm (see figure 1.4).

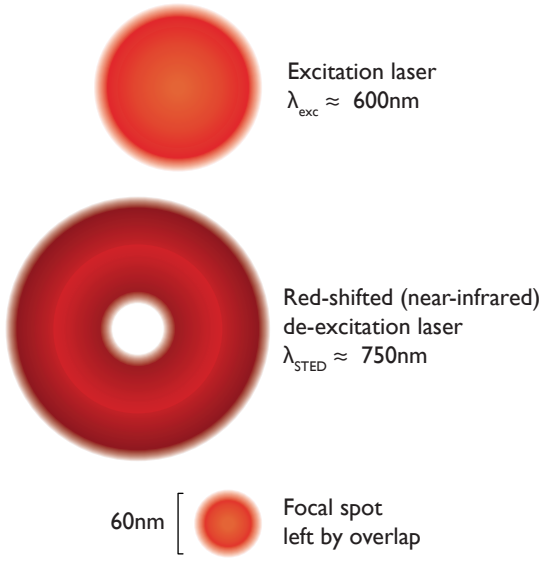


Figure 1.4. Lasers in STED microscopy
STED microscopy achieves super-resolution images by using two collinear laser beams to excite molecules in a very small area, allowing individual molecules to be illuminated separately. This is achieved with two lasers, one that excites the fluorophores, and another doughnut-shaped de-excitation laser. These are fired in very rapid overlapping pulses, resulting in just those molecules in the doughnut hole being able to fluoresce. Point spread functions are then mapped to the well-discriminated fluorescent events, and these used to precisely determine the location of the fluorescing molecule.

In contrast, Stochastic Optical Reconstruction Microscopy[156] (STORM) discriminates between molecules by using a weak light source to excite a random selection of molecules in the sample. Provided each fluorescing molecule is separated by a distance equal to no more than half the required resolution, they can be distinguished accurately from their point spread functions. The beauty of STORM is that, in its original form, only a single red laser is needed to both turn on and turn

[iii] In 1952, Schrödinger wrote: 'we are not experimenting with single particles, any more than we can raise Ichthyosauria in the zoo'[159]. Which goes to show that Jurassic Park is no laughing matter.

off the Alexa Fluor dye. This means that a steady application of low intensity light will cause the labelled molecules to continually blink on and off, allowing a picture to be built up gradually over the course of a few minutes. Though not as fast as other super-resolution methods, STORM, PALM (photoactivated localisation microscopy) and others in this category can achieve incredibly good resolutions of just a few nanometres in scale.

Super-resolution protein complexes

Most of the complexes studied to date with SMLM have been large and highly abundant, thus lending themselves well to the technique. Ground state depletion microscopy for example (which is related to STED) has been used in particularly attractive work determining the relative location of different Nup subunits in the nuclear pore complex[160]. This was achieved using a system coupling GFP-Nup fusion proteins with fluorescent anti-GFP nanobodies to localise each subunit[161]. Making use of image classification in a similar fashion to cryo-EM, multiple NPC particles were averaged to generate highly accurate measurements of the diameter of the pore formed by each subunit. They were then able to determine the relative location of each subunit within the larger complex by comparing the diameter of these rings.

1.5 COMPUTATIONAL PREDICTION OF PROTEIN COMPLEX STRUCTURE

Frequently, experimental methods for determining protein complex structure are not possible, or tellingly, are simply no longer the best use of a researcher's time and money. Prediction of protein structure from sequence, first prophesied by Anfinsen in 1973[162], has been a long-standing challenge that has until recently been impossible in practice, for want of both sequence data and computational power. However, the genomic era has seen exponential increases in both of these areas, along with a similar expansion in the number of experimentally determined protein structures. Accordingly, there has been a concomitant improvement in our ability to predict structures computationally. Unusually for academia, two important and long-running competitions have been instrumental in driving progress in the field. Both competitions - CASP (Critical Assessment of protein Structure Prediction[163]) and CAPRI (Critical Assessment of PRediction of Interactions[164]) - are now an integral part of the community, providing much-needed benchmarks and progress reports that enable both users and developers to keep track of the rapidly changing state of the art.

Broadly speaking, the field of protein structure modelling can be divided into two overlapping subgroups, albeit separated mostly by their philosophical stances. Older, and currently more practical, are top-down approaches based on the use of templates for the structures being modelled; these templates are selected based on close sequence similarity with the target protein or complex. In contrast, there is equal interest in prediction of structure from first principles - an approach exemplified by molecular dynamics (MD). Template-based modelling (TBM) and MD represent opposite sides of the protein modelling community, but in practice there is a great degree of overlap between the two, and most of the methods that I will describe below use elements from both.

1.5.1 Top-down modelling of protein complex structure

The extent of the improvement in predictive power is such that, for individual sequences with close sequence similarity (60+%) to known structures, it is usually possible to produce structures that are within a few ångströms of the experimentally determined version, as measured by root-mean-square-deviation of residue distances and other metrics[165,166]. This is the process known as TBM, and nowadays it is routinely used to facilitate experimental structure determination of single protein chains.

For protein complexes, regardless of whether the structures of individual subunits are already known, it is often possible to reach a realistic approximation of the correct complex structure by a combination of homology modelling and molecular docking. Though most applicable for investigating protein-ligand interactions, molecular docking combined with homology modelling is becoming increasingly viable for protein-protein interactions, as evidenced by numerous recent studies and the results from the CASP and CAPRI competitions[167–170].

Template-based modelling

TBM is based on the principle that the degree of sequence divergence in homologous proteins is closely related to their structural similarity[171]. Once a suitable template is found - realistically with at least 40% sequence identity with the target protein - the sequences are aligned and conserved regions are used to map fragments of the target onto the template structure. This is followed by replacement of the loop regions and final model refinement.

There are numerous methods based on extensions of this basic protocol that enable modelling of complete protein complexes, in addition to individual subunits[172–174]. One important and widely used strategy for template identification is threading, or dimeric threading, in the case of modelling complexes[175,176]. Threading differs slightly from approaches based solely on sequence homology, in that it relies more on fold recognition than sequence similarity - this is assessed by a scoring function - the template that is eventually selected is the one which minimises this function. As a general rule, threading is used when the target sequence has particularly low sequence similarity with other known proteins, but in practice, most modern software takes these decisions out of the hands of the user. For a broad overview of the currently available software and experimental strategies for TBM, readers should see the recent review on the topic by Szilagyi and Zhang[177].

Prediction of protein-protein interfaces

Template-based methods work by mapping the sequence of the target proteins of interest onto a template of the protein complex, without explicitly modelling the interface until later refinement steps. In contrast, 'molecular docking' begins with subunits in their monomeric form, and models the interface directly by attempting to minimise the potential energy landscape of the bound proteins. This is achieved by sampling the conformational space of the two proteins with respect to each other and scoring the different interfaces that can be formed between them[178].

These two steps (sampling and scoring) can be handled in a number of ways. Conformational sampling is computationally very intensive, since two chains moving in three dimensions produces

six degrees of freedom from the get-go, and allowing side-chain and backbone movements increases this number drastically. Thus, at least initially, almost all docking methods assume the proteins of interest to be rigid bodies, and fix the orientation of one protein with respect to the other. This reduces the complexity of the sampling problem greatly. Fast Fourier transforms[179] were the first method to make molecular docking possible, but other approaches have subsequently been developed too. These include Monte Carlo search[180] and normal mode analysis[181], the former of which is of note because of its use in RosettaDock[182,183] - one of the most popular and successful docking programs. A second popular program is HADDOCK[184,185], which uses a gradient-based search method.

Scoring of interfaces is another non-trivial problem. As in the wider field of structure modelling (which encompasses docking) solutions to this problem tend to take the form of either physical or empirical, knowledge-based methods. In the former camp, force field scoring functions[186] are used to model the energy of the system in a given conformation, and typically involve a large number of parameters relating to attributes such as Van der Waals interactions and intramolecular strain energies. The latter consists of conceptually simpler techniques including counting of intermolecular contacts and scoring based on prior knowledge of statistically likely interactions gleaned from sources such as the PDB[187,188].

1.5.2 De novo structure prediction

The docking methods described above are contingent on having structures available for the proteins whose interactions you are trying to model. However, it is often the case that there is no solved structure or suitable template available for use with homology modelling. In the past, this would have meant that the best one could do would be to try and predict secondary structure regions and likely interfaces or infer the presence of binding domains from homologous sequences using tools such as JPred4[189] or databases such as PFAM and UniProt[190,191]. With this in mind, researchers have begun to make inroads into de novo structure prediction, as well as looking at ways in which the difficulties of true de novo prediction can be circumvented using sequence information alone.

Using protein coevolution to infer intermolecular contacts

Using the evolutionary sequence record to inform structure prediction is an idea that has been around for at least 20 years[192], but has been held back by the fact that it is very challenging to distinguish coevolving sites that indicate direct amino acid contacts from transitive ones. For example: if residues A and B are in contact, and residues B and C are in contact, then A and C may also show a strong coevolutionary signal, despite interacting via an intermediary residue. Over the entire protein sequence, this blurring of coevolutionary signal is sufficient to prevent meaningful structure prediction. The major breakthrough in tackling this problem was achieved with the development of an algorithm named Direct Coupling Analysis[193,194], which extends Shannon's concept of mutual information[195] and enables direct and indirect residue contacts to be distinguished from each other. Deborah Marks and her colleagues have now implemented this algorithm in a more generally applicable and user-friendly format, enabling the method's widespread use in the structure-prediction

community[196–198].

Though originally used for single protein structures, this method is equally applicable to protein complexes, as intermolecular contacts are subject to many of the same coevolutionary pressures as intramolecular ones. EVcouplings (from the Marks group) has recently been applied to protein complexes[198]. Out of a set of 82 protein complexes with unsolved structures, 32 had a sufficiently good sequence record as to be able to predict the entire complex de novo, whereas others were sufficient to predict intermolecular contacts, but not the entire structure. Unfortunately, a limitation of this technique is identification of homomeric contacts, since without additional information these cannot be distinguished from intramolecular interactions. A related issue is that many nominally heteromeric interactions arise from homomeric interactions between genes that have undergone duplication and subsequent genetic drift[199–201]. In such cases, it may not be possible to acquire a sufficient number of sequences for structure/interaction prediction, particularly if the proteins in question have diverged recently.

Molecular Dynamics

At present, we are still a long way from saturation of structure and sequence space, as is clear from recent studies sampling viral and prokaryotic genomes[202–204]. As such, a substantial proportion of the protein universe is beyond the reach of either TBM or EVCouplings. Molecular dynamics is a simulation method that models the behaviour of all the atoms or molecules in a dynamical system. This is typically achieved, as for the force field method introduced earlier, by numerically solving Newton's equations of motion for the entire system of particles. Being rooted in basic physical principles, and unlike most of the other techniques we have discussed, MD is enormously versatile, and is used widely in other fields outside of biology[205–207]. From our perspective, much of its power lies in its ability to span multiple cellular scales - it has been used to study processes ranging from the molecular details of ligand binding to the assembly of virus capsids[208,209].

Though ab initio MD methods exist that explicitly include the electronic configurations of all atoms as parameters in the model, these are not practical for systems beyond a few atoms in size and across timescales of more than a few picoseconds. For biological macromolecules on the scale of protein complexes, empirical force fields are used instead. These describe an approximation of the potential energy of the system through a function with the general form $U(R)$, where $R = \{r_1, ..., r_n\}$ describes the coordinates of the atoms in the system. It is this same potential energy function that is minimised in molecular docking simulations. In slightly more detail, $U(R)$ can be described as follows:

$$
\begin{aligned}
U &= E_{bonding} + E_{non-bonding} \\
E_{bonding} &= E_{bonds} + E_{angles} + E_{torsions} \\
E_{non-bonding} &= E_{electrostatic} + E_{Van\ der\ Waals}
\end{aligned}
$$

Each of the energy terms in the above equations are themselves functions of the atom coordi-

nates and empirically determined constraints such as bond strength, molecular weight and so on. Though computationally more efficient for ignoring the electronic degrees of freedom, a drawback of force field methods is that large changes such as the making or breaking of covalent bonds cannot be modelled. However, in the case of protein complexes this is rarely a problem. Many different force fields have been tailored to different purposes, but for studying protein-protein interactions AMBER[36,37] and CHARMM[35] are particularly relevant. The dedicated structure prediction software FoldX[210] uses a proprietary force field, though with similar construction to the general one described above.

At present, computational power is only just beginning to reach the level at which the assembly of protein complexes can be modelled from scratch. The two major obstacles barring further progress are the size of proteins, and the timescales on which assembly takes place. Currently, only small proteins or protomers can be modelled, and even then only across very short timescales. This is unfortunate, since folding and protein complex assembly takes place over timescales of microseconds to minutes. However, a recent exciting study by Plattner and colleagues[211] has had some success using hidden Markov models in combination with thousands of MD simulations starting from different points, selected so as to maximise the efficiency of conformational space exploration. This enabled them to model the bacterial ribonuclease barnase in complex with its inhibitor barstar with impressive accuracy. Associated, dissociated and transition states were observed across 2 milliseconds of modelling time, and predicted thermodynamic parameters were in strong agreement with those obtained from independent experimental results.

1.5.3 Protein complex databases and repositories

Since the first print-format directories of protein sequences were produced by Margaret Dayhoff[32], protein sequence and structure databases have become an integral part of the research infrastructure in biology. These databases range in size from manually curated lists of proteins or protein complexes involved in specific cellular processes, to vast data repositories such as UniProt or the PDB. The benefit of these larger repositories is clear, since without them the task of collating data from different experiments would be impossible for any single researcher. However, smaller datasets from individual papers are often extremely useful too, as they offer nuance and context that can't be obtained from data compiled from multiple sources. A selection of useful databases and repositories of protein complexes or interactions are summarised in table 1.1 below.

1.6 ANSWERING QUESTIONS ABOUT THE PROPERTIES OF PROTEIN COMPLEXES

Armed with the methods covered in the preceding pages, there are numerous biological questions we can attempt to answer about the cell and the behaviour of protein complexes within it. Most obviously, the question of what proteins look like is enabled by structural methods such as X-ray crystallography and cryo-EM. This is useful beyond simple curiosity, as understanding the mechanistic details of a protein or protein complex's action enables us to make better decisions about the design of future experiments or drug development. More broadly, we can also investigate how the cell maintains protein stoichiometry, and how these mechanisms interact with and facilitate the

Name	Description	References
Protein Data Bank	Very large repository of structural information on proteins and protein complexes.	[58]
IntAct	IntAct Molecular Interaction Database. Manually curated database of protein-protein interactions compiled from literature sources and direct user submissions.	[212]
Complex portal	Manually curated list of protein complexes from a selection of model organisms. Related to IntAct.	[213]
hu.MAP	Human Protein Complex Map. Generated by integrating three previously released datasets using machine learning.	[105,106,130,214]
CORUM	The comprehensive resource of mammalian protein complexes. Starting to show its age somewhat, but still widely used as a gold standard set of experimentally validated mammalian protein complexes.	[215]
Complex census	A Census of Human Soluble Protein Complexes. This is an important paper that was released in 2012, covering 3000 human soluble complexes. Data was obtained by biochemical fractionation and MS.	[216]
BioPlex 2.0	AP-MS study of the human interactome, covering some 25% of the human proteome.	[130,217]

Table 1.1.: Useful repositories for research on protein complexes

assembly of individual complexes. These two questions are the primary focus of this thesis.

1.6.1 Maintenance of cellular stoichiometry

A typical cell contains thousands of different protein species - less in most bacteria and archaea, and more in most eukaryotes, particularly multicellular ones such as ourselves. Imbalances in the expression of proteins are usually deleterious, and a variety of aberrant behaviours can be attributed to them; cancer, for example, can be caused by changes in the expression of various oncogenes or tumour suppressors.

More generally, members of heteromeric protein complexes are particularly sensitive to changes in expression, as they impact not just on the protein in question, but also those that it interacts with. Strong evidence for this comes from observations that yeast strains that are heterozygous for single gene knockouts tend be particularly unfit when those genes are members of complexes[46]; similarly, mass spectrometric studies[218,219] (using iBAQ and SILAC, respectively) have shown that ribosomal subunits produced in excess are rapidly degraded by the ubiquitin proteasome system. This is telling, since it implies that there are significant fitness costs to having unbound subunits free in solution indefinitely. There are many reasons why this might be, including the need to avoid protein mis-interactions, which imposes considerable constraints on the evolution of protein surfaces[220].

Surprisingly however, early pulse-chase studies on the assembly of alpha and beta spectrin in birds and mammals showed that these subunits were not being synthesised in stoichiometric proportions,

but rather in highly unequal ratios, with excess protein being degraded later[221–223]. From an anthropomorphic design perspective this is unexpected, since it would presumably be more efficient to avoid imbalances in the first place.

To what extent can the relative synthesis rates of subunits be controlled? Is spectrin a special case or is unequal synthesis and degradation the norm? One would naïvely expect the most efficient approach to be production of subunits precisely in accordance with their stoichiometry, but measuring this accurately is challenging. This is because of the different ways in which gene expression can be controlled, namely synthesis and degradation rates at both mRNA and protein levels. Mass-spectrometric analyses of whole-cell protein extracts give an idea of steady state protein abundances, but it is difficult to deconvolve the relative contributions of synthesis and degradation rates. Metabolic pulse-chase experiments such as those used for spectrin have had more recent successes when used in combination with SILAC MS[224], but these are still not perfect due to the lag times incurred by the labelling process.

An elegant method for measuring translation rates directly exists in the form of ribosome footprint profiling[225,226]. This method works by deep sequencing ribosome-protected mRNA fragments and counting the number of fragments associated with each gene. The assumption underlying this being that every ribosome bound to an mRNA will produce one protein, and therefore that the number of ribosome-protected fragments is a good indicator of translation rate.

Ribosome profiling has been used to demonstrate that, in *E. coli* at least, stoichiometric synthesis is far more common than unequal synthesis[227]. This is less surprising than would be the case in eukaryotes, since many protein complexes are encoded within operons, which likely removes much of the potential variance in expression due to transcription. However, even for complexes such as F_0F_1 ATP synthase, in which some subunits are present in many copies, subunits are translated in the correct proportions, with significant differences in ribosome density for different genes within the polycistronic mRNA.

Less clear is whether proportional synthesis is also the norm in eukaryotes. Limited ribosome profiling data from *S. cereviasae* suggests that some complexes at least are synthesised stoichiometrically[228], but the majority of studies (including the work described in chapter 3) suggests that most eukaryotes do not show the same tight control of synthesis seen in bacteria.

1.6.2 Assembly of protein complexes

Another theme running throughout this thesis is the importance of ordered assembly of protein complexes. A method pioneered by Carol Robinson's group in Oxford was nano-ESI MS, which allowed dynamic processes such as the assembly or disassembly of protein complexes to be observed in real-time[114]. The first paper to make use of this method explicitly to study the assembly order of protein complexes focused on homomers[103], and demonstrated that both assembly order and the evolutionary trajectory of subunit gain could be predicted well by interface size.

This work was subsequently extended to heteromers, showing that, like homomers, assembly pathways of heteromers can also be predicted by a hierarchy of interface sizes[104]. These are also evolutionarily conserved, with gene fusions being significantly more likely to occur in cases in which

assembly order is preserved. More recently, it has also been shown that the number of coevolving residues between two proteins is a similarly good predictor of assembly order[229].

The natural question arising from the ability to determine assembly pathways in vitro is how much they matter in vivo? Evidence of evolutionary conservation is a good indicator of biological importance, but to what degree does efficient assembly matter to eukaryotes compared to prokaryotes? One way in which ordered assembly could be facilitated is through co-translational assembly of protein complexes. In prokaryotes, there is direct evidence for this from a study on the assembly of luciferase, which made use of fluorescence resonance energy transfer to show that the molecule was assembled most efficiently co-translationally from subunits encoded on the same operon[44]. In eukaryotes, despite the very different genomic organisation and scale on which assembly takes place, there is nonetheless good evidence for co-translational assembly being common[230], further supporting the idea that efficient assembly is universally important and therefore worth continued investigation.

1.7 DISCUSSION

The study of protein complexes is currently undergoing a sea-change, brought about by the recent breakthroughs in structural biology, the emergence of mass spectrometry as a quantitative tool, and ongoing developments in computational techniques. The methods presented here offer a broad selection of those that can be used to study the physical characteristics and behaviour of protein complexes, but unavoidably there are some omissions, such as small-angle X-ray scattering[231] and hydrogen-deuterium exchange MS[232].

A common theme that I have tried to highlight in this review is the overlap between the many different fields concerned with characterising protein complexes. Most of these overlaps have had a synergistic effect on the technologies involved; this has been particularly obvious in cryo-EM, where hardware improvements have directly driven the development of new image processing software, but many other examples exist across structural biology and further afield. To point out a few explicitly: homology modelling has enabled much faster processing of diffraction patterns and electron density maps, improvements in purification techniques benefit essentially all of the non-computational techniques we have discussed, and many of the advances in imaging in cryo-EM will likely be transferable to XFELs.

This leaking of technologies across fields has facilitated the rise of integrative structural biology, which is becoming the most powerful approach to investigations on protein complexes. Many of the most impressive structures published in the past couple of years have been the product of combinations of methods, for example: the transcribing mammalian PolII complex[233], the nuclear pore complex mRNA export platform[234], and the Mediator complex[235]. A second common feature of all of these papers is their focus on mechanistic descriptions of function or assembly, demonstrating a welcome move away from purely descriptive studies. Given the direction the field is moving in, early-career structural biologists should not to be content with specialising in one method or the other[51,236], but should endeavour to be at least familiar with most of the topics covered here.

The shift from purely descriptive studies to mechanistic ones emphasises the fact that there is

more to proteins and protein complexes than simple descriptions of structure. Of particular importance for the rest of this thesis, there is much to be gained from understanding the assembly process of protein complexes. Indeed, thanks to native MS studies, it is now well established that this occurs along ordered, thermodynamically favourable pathways, and papers on the topic have been published continually since this fact was first demonstrated[103,104,229,237,238].

On a larger scale, inventive use of mass spectrometry is enabling rapid improvement in our understanding of how individual protein complexes fit into the wider proteome. In a standout study from the group of Matthias Mann[106], the proteome of HeLa cells was quantified in such a way as to accurately capture interaction stoichiometries and global cellular abundances. Although not unexpected, the results from this work clearly demonstrate that the large majority of interactions, though important, are fairly weak. In contrast, stable complexes formed from interactions with stoichiometric ratios on the order of 1:1 are significantly rarer, but nonetheless highly connected through these weaker interactions.

The long-term objective of the techniques we have discussed in this review is to give a complete and unified understanding of the cellular proteome, in both its constituent parts and its behaviour at scale. The progress we have made towards this aim would scarcely have been imaginable to the researchers who first began studying proteins in the 1950s, there is no reason to suspect that the next 50 years will not see even greater progress. In many of the fields I have discussed there are novel technologies that will be revolutionary in years to come - nowhere more so than with the development of XFELs and serial femtosecond X-ray crystallography; it will be fascinating to see the new studies that this technology enables. Alternatively, perhaps cryo-EM will be able to continue along its current trajectory to overtake crystallography as the go-to method in structural biology?

In the field of mass spectrometry, although there are no obviously disruptive technologies on the horizon, continuing improvements in the sensitivity and accuracy of detectors are assured. Algorithmic development in MS is another area in which improvements are needed. Currently, there are seemingly intractable issues with peptide discrimination and quantification that need addressing, as evidenced by our first serious attempts to map the human proteome[239–241]. Nonetheless, the inherent versatility of the method across different cellular scales ensures the field's relevance in the decades to come, particularly as we move towards single-cell biology[242].

Last, but by no means least, computational modelling of proteins continues to go from strength to strength. Though not a panacea for the difficult challenges we still face in structural biology, as the diversity of sequences and structures increases, so too will our ability to leverage computational power to fill in the gaps through homology modelling. Molecular dynamics too is finally approaching a point at which we can use it to study real-time processes of complex assembly, and with tantalising hints of progress in the world of quantum computing, the future looks bright for the study of protein complexes.

2 OPERON GENE ORDER IS OPTIMISED FOR ORDERED ASSEMBLY OF PROTEIN COMPLEXES

2.1 INTRODUCTION

Work carried out over the course of the last decade has revealed that protein complexes, both homomeric and heteromeric, assemble via ordered, energetically favourable pathways. Assembly and disassembly pathways can be observed in vitro using ESI-MS[114], and by looking at gene fusions it has been shown that these pathways are evolutionarily conserved[103,104]. These experiments are laborious and time-consuming, but fortunately assembly order can be predicted with good accuracy computationally if the structure of the complex is available; in most cases, assembly order is simply determined by interface size, with larger interfaces assembling earlier. This has been confirmed independently by a recent study using a combination of MS, NMR spectroscopy and EM, and a second study showing assembly order can be predicted by the number of coevolving residues between different subunits[229,238].

Given the central importance of protein complexes to most biological processes, there is a strong pressure on the cell to ensure that they assemble correctly in a timely and efficient manner. However, this process is inherently stochastic, and takes place in a cellular environment in which the background concentration of protein and other biological macromolecules is incredibly high[243]. For heteromeric protein complexes, this presents a serious problem: how do protein subunits expressed from different genes find each other in such a crowded environment? The cost of failure is not trivial, since misassembly or non-specific interactions with other proteins can lead to the formation of toxic aggregates. To give an example familiar to many of us, the formation of such aggregates is implicated in a number of neurodegenerative diseases[244].

If the assembly pathways that are observed in vitro also occur in vivo, then we might expect to see evidence of this in the regulatory systems possessed by the cell. In bacteria and archaea, protein complexes are often encoded within operons[245,246], where multiple genes are transcribed onto a single polycistronic mRNA. This presents a possible opportunity for enhancing the efficiency of protein complex assembly - if physically interacting genes are closer together within operons, then this would increase the likelihood of those subunits finding each other upon being translated. Thus, we reasoned that there might be some correspondence between operon gene order and assembly order. The results from this work show that this is indeed the case.

2.2 RESULTS

2.2.1 Encoding protein complexes within operons is likely to facilitate efficient assembly

In order to test the hypothesis, we acquired a large set of heteromeric protein complex structures from 70 bacterial and archaeal species. Describing these protein complexes as lists of non-redundant gene/subunit pairs, we then mapped the location of each gene pair in the genome of the species it came from. Out of a total of 1079 gene pairs, 368 were encoded within the same transcriptional unit - that is, translated from the same mRNA (figure 2.1A), with the remaining 711 being transcribed separately.

By transcribing genes in operons, any differences in expression levels due to transcription rate are automatically removed, and instead variation in observed protein abundances must be due to differences in translation or degradation rates. It has been suggested therefore that one of the primary benefits of encoding protein complexes in operons is the reduced stochasticity in gene expression associated with operons[44,247,248], which is consistent with observations demonstrating that stoichiometry of most protein complexes is tightly controlled in *E. coli*[227]. In figure 2.1B, we demonstrate that in *E. coli*, as expected, protein abundances (obtained from PaxDB[249]) of gene pairs encoded in the same transcriptional unit are more closely correlated than those not. The same trend was seen when combining data across all organisms for which structures and operons were available, as well as when using absolute protein synthesis rates using ribosome-profiling data[227] (figure A.1).

The likelihood of protein complex subunits randomly encountering each other in the cell is greater for highly expressed complexes. Since operons will necessarily lead to co-localisation of its freshly translated proteins, the benefit to being operon-encoded should be particularly strong for lowly expressed protein complexes[243,250]. Supporting this prediction, in figure 2.1C, we show that there is a highly significant tendency for operon-encoded subunits to be less abundant, with an approximately order of magnitude difference in the median abundance of the two groups.

2.2.2 Adjacent genes within operons are more likely to physically interact

As suggested by the fact that lowly expressed protein complexes are more likely to be encoded in operons, close proximity upon translation probably enhances the efficiency of assembly (figure 2.2A). Supporting this idea, others have noted previously that adjacent genes are more likely to physically interact[245,246], though these studies do not explicitly focus on operon-encoded genes.

When comparing the number of adjacent genes that physically interact (208 interacting pairs out of 220 total) with the number of non-adjacent interacting pairs (77 out of 148), there is a much higher tendency for adjacent genes to share a physical interface (odds ratio = 15.8, p-value = 5×10^{-22}, Fisher's exact test). In figure 2.2B we show that this tendency extends beyond just adjacent genes, with the effect continuing for the first two intervening genes. A highly similar trend is observed using pairwise interaction data obtained from a large Y2H screen in *E. coli*[28] (figure 2.2C); this is again significant when comparing the likelihood of adjacent vs. non-adjacent gene pair interaction (odds ratio = 2.7, p-value = 0.0002), despite the apparent weaker ability of

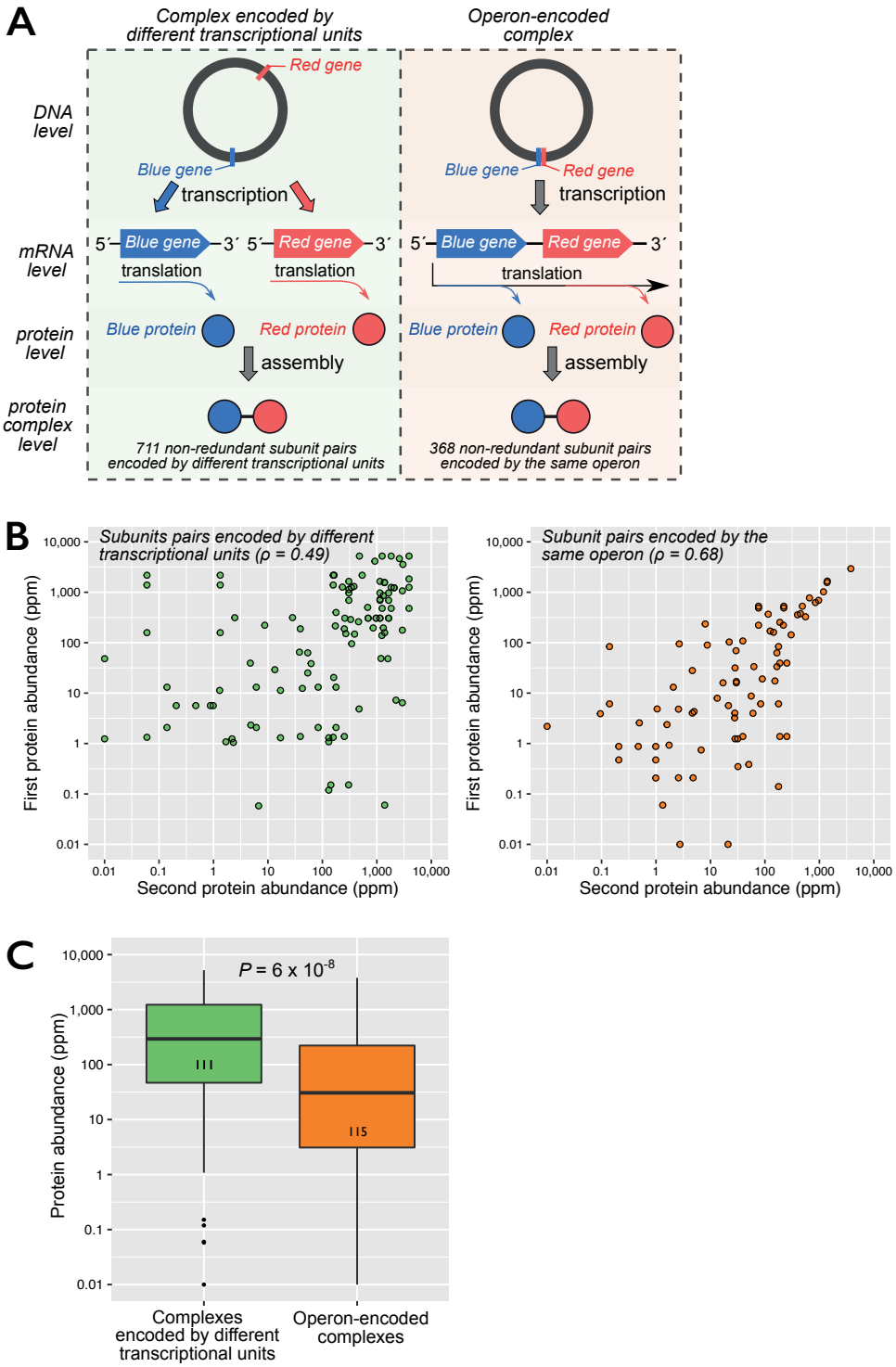


Figure 2.1.: Encoding protein complexes within operons enhances assembly efficiency
(A) Transcription, translation and assembly for heterodimeric subunits encoded by the same vs. different transcriptional units. (B) Differences in protein abundance correlations (Spearman's ρ) for gene pairs encoded within the same vs. different transcriptional units. The correlation between genes encoded within the same operon is significantly higher than for those in different transcriptional units (p-value = 0.002), as determined by randomly shuffling pairs between groups 10^5 times. (C) Protein complexes encoded within operons are significantly less abundant on average than those encoded in different transcriptional units, with significance determined using the Wilcoxon rank-sum test. Adapted from figure 1, Wells et al.[1]

Y2H to detect interactions.

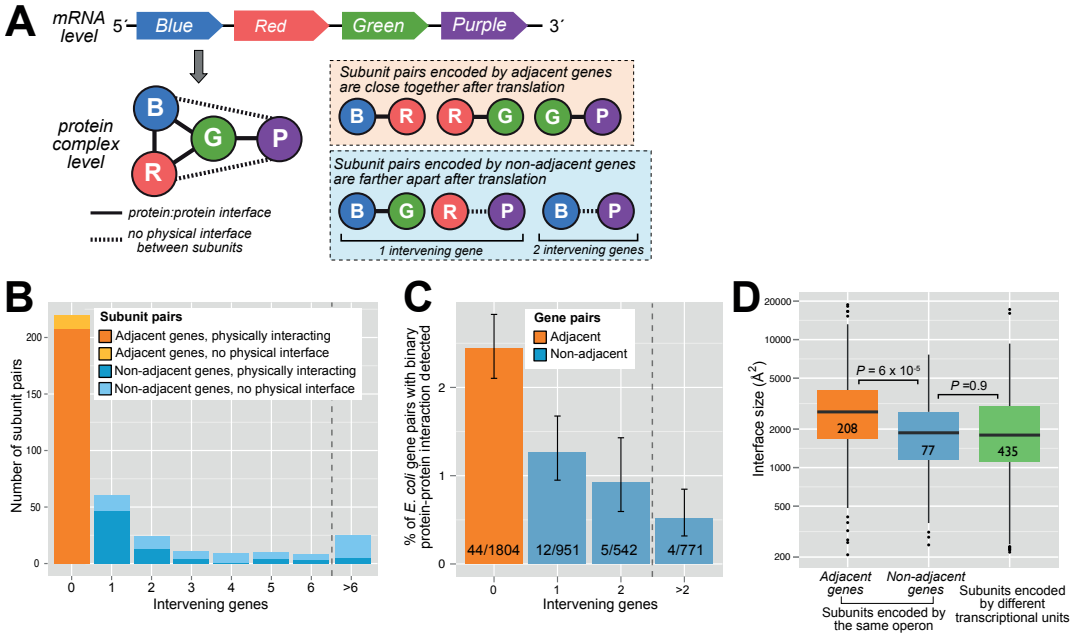


Figure 2.2.: Adjacent genes within operons are more likely to encode physically interacting subunits
(A) Within a given complex, although all subunits interact indirectly, not all must necessarily share a physical interface. Within operons, the genes that code for these subunits can either be adjacent or non-adjacent. (B) Subunit pairs separated by number of intervening genes. Each bar is subdivided into those pairs that physically interact and those that don't. We define a physical interaction between two genes as their sharing an interface of > 200Å. (C) Analogous to B, but using binary interaction data obtained from Y2H screens[28]. Error bars are 68% Wilson binomial confidence intervals. (D) Physical interfaces between adjacent genes are significantly larger on average than either non-adjacent genes or those encoded in different transcriptional units. P-values were calculated using Wilcoxon rank-sum tests. Adapted from figure 2, Wells et al.[1]

Since the median length of operons in our dataset is fairly small (four genes), the number of possible gene pairs is only slightly skewed towards non-adjacent: a four gene operon has three adjacent and three non-adjacent pairs, but the number of non-adjacent pairs increases rapidly as operons get larger. We therefore reasoned that the apparent tendency of adjacent gene pairs to physically interact might just be an artefact of the high representation of such pairs in our dataset. To control for this possibility, we generated a null model in which all of the genes were shuffled within their operons. We then compared the number of observed gene pairs that were both adjacent and interacted with the number expected under the null model, and repeated this process 10^5 times (figure 2.3A). In all cases the number of observed interacting pairs was significantly greater than expected by chance. As before, this is also significant, to a lesser degree, when using pairwise interaction data generated by Y2H assays with *E. coli* (figure 2.3B).

Since many of the possible gene pairs in our dataset arise from a particularly large operon in *Thermus thermophilus* (the *nqo* operon, which encodes respiratory chain complex 1), we repeated these analyses excluding this operon, and considering this operon alone (figure A.2). These were both significant, indicating that the result was not due to features unique to this operon. It also demonstrates that the trend holds within a single operon, provided that it is sufficiently large.

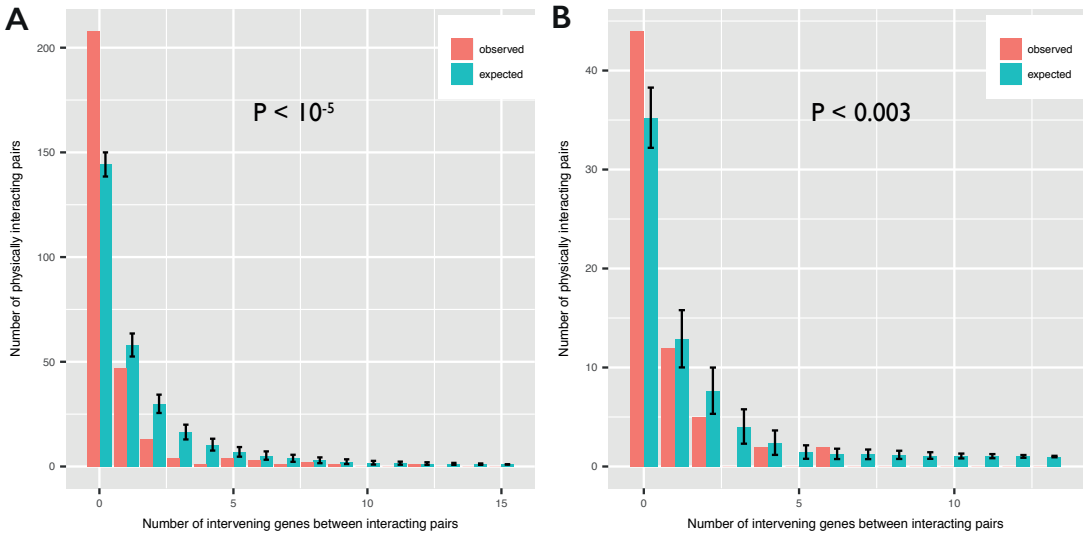


Figure 2.3.: Relationship between gene pair proximity and likelihood of physical interaction
(A) The number of physically interacting genes at different distances compared to a null model in which gene order in operons is shuffled. (B) shows the same result using binary interaction data from Y2H screens[28]. To assess the significance of the observed tendency for genes closer together within an operon to physically interact, permutation tests were performed by shuffling the order of genes within operons 10^5 times. In each trial (i.e. a single shuffling of gene orders), the total number of intervening genes between physically interacting subunits was counted; to calculate the p-value, the number of occasions that this expected number of intervening genes (based on shuffled operons) was less than the true figure was divided by the number of trials. Adapted from figure S2, Wells et al.[1] See also figure A.2.

The observation that adjacent genes are more likely to physically interact could have an effect on the interpretations of earlier work showing that gene fusion events tend to preserved the order of assembly[104], since adjacent genes often undergo fusion events[251]. To test whether this earlier observation was affected by our newer findings, we repeated the test for assembly-conserving fusions using only adjacent genes, and found that there was still a significant tendency for fusions to conserve assembly order (figure A.3).

In addition to the increased tendency of adjacent gene pairs within operons to form a physical interface, we also observed that these interfaces are typically larger than those formed by non-adjacent gene pairs. Figure 2.2D shows the distribution of interface sizes for those proteins that physically interact within a protein complex. Whilst there is significant overlap in the interface size distributions (since they are not normalised between complexes), there is nonetheless a clear difference between adjacent and non-adjacent gene pairs, regardless of whether encoded within the same operon or not. This is surprising, and hints at a relationship between gene order and assembly order, since we know that larger interfaces tend to form earlier than smaller ones[103,104,238].

2.2.3 Operon gene order closely matches order of assembly

Building on this finding, we next considered whether or not there could be a correlation between the assembly order of protein complexes and their gene order within operons. Thus far, we have demonstrated that there is a spatial relationship between gene order and protein complex assembly, in that adjacent genes are more likely to physically interact, and form larger interfaces when they do

so. However, since the assembly of a protein complex does not occur instantaneously, there is also a temporal aspect of the problem to be considered. Within bacterial and archaeal systems, there are two factors that impose a temporal order on the expression of genes encoded within operons. The first of these is the phenomenon of co-transcriptional translation, in which ribosomes begin translating nascent mRNA as it is being transcribed[252–254]. The second is translational coupling, in which translating ribosomes proceed directly from one gene to the next, without being released from the mRNA[255,256].

Both of these have the effect, initially at least, of ensuring that genes towards the 5' end of the mRNA transcript will be translated before those at the 3' end. Thus, if protein complex subunits that tend to assemble earlier were encoded at the start of operons, then this would likely increase the efficiency of assembly. There are three different ways in which a given pair of adjacent genes can assemble once translated. If the two proteins form a heteromeric interface, then assembly is simultaneous and gene order is interchangeable for that pair without affecting assembly efficiency. Alternatively, if the first interface to form is homomeric, then gene order does have an effect - either gene order can match assembly order, in which case the first gene to be translated will also form an interface first, or it can be different from assembly order. In figure 2.4A we demonstrate these three different scenarios.

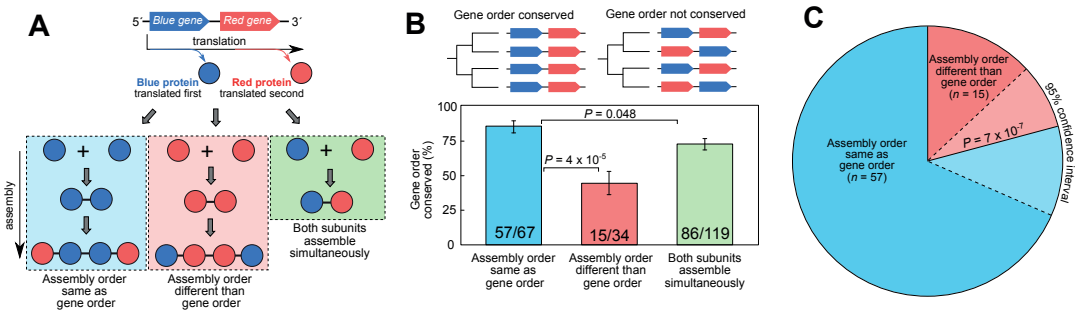


Figure 2.4.: Operon gene order reflects protein complex assembly order
(A) Three possible scenarios for the relationship between gene order and assembly order for a single gene pair. (B) Conservation of gene order for the three possible relationships described in A. Error bars are Wilson binomial confidence intervals and p-values were calculated with Fisher's exact test. (C) Assembly order matches gene order in 79.2% of gene pairs whose order is evolutionarily conserved. P-value was calculated using the binomial test. Adapted from figure 3, Wells et al.[1]

To investigate the potential relationship between gene order and assembly order, we predicted assembly pathways for all of the operon-encoded heteromeric gene pairs in our dataset, then separated all of the 220 adjacent pairs into the three categories described above. We then considered the tendency of gene order in each of these three groups to be evolutionarily conserved. As shown in figure 2.4B, there is a significant tendency for gene order to be conserved in cases where it matches assembly order. This suggests that gene order is constrained by the requirement that it not interfere with protein complex assembly.

Considering the 72 pairs where gene order is evolutionarily conserved and assembly of one subunit happens before the other, we found that there was a striking correspondence between gene order and assembly order (figure 2.4 C). In 57 out of 72 pairs gene order matched assembly order

(79%, p-value = 7×10^{-7}, binomial test), compared to just 10 out of 29 cases (34%) where gene order was not conserved. Thus, selection for ordered protein complex assembly appears to be a major driver of gene order in prokaryotic operons.

The likelihood of a physical interaction between two protein decreases as the number of intervening genes between them on an operon increases (figure 2.2). We see a similar trend when looking at the relationship between gene and assembly order, with the two features becoming less likely to match as the distance between the two genes on the operon increases (figure A.4). This is unsurprising, as once proteins have dispersed around the site of translation, the beneficial effect of protein co-localisation arising from tightly controlled gene order becomes lost.

A feature of operons that some studies have noted is for genes towards the start of operons to be expressed at higher levels[257,258] [i]. This leads to a possible alternative explanation for the correspondence between gene order and assembly order. If there is some requirement for earlier assembling proteins to be expressed at higher levels, then this could produce the same observation for different reasons. To rule out this hypothesis, we note that there is no relationship between assembly order and protein abundance and that gene order is a better predictor of assembly order than abundance (figure A.5). We do however see a weak, though insignificant, trend for upstream genes within operons to be more abundant (table 2.1).

	Subunit encoded by upstream gene is more abundant (*E. coli* only)	**Subunit encoded by upstream gene is more abundant (All species)**	**Subunit encoded by upstream gene has a higher rate of protein synthesis**
Gene order is evolutionarily conserved	26/47 (55.3%)	36/67 (53.7%)	20/43 (46.5%)
Gene order is not evolutionarily conserved	5/10 (50.0%)	10/16 (62.5%)	3/9 (33.3%)
Total	31/57 (54.5%)	46/83 (55.4%)	23/42 (44.2%)

Table 2.1.: Relationship between gene order and abundance for adjacent heteromeric subunits
Previous works have reported a significant tendency for upstream genes in operons to be more abundant[257,258]. Looking only at proteins that are members of protein complexes and encoded adjacent to each other within operons, we see a slight tendency for upstream genes to be more abundant. However, this is not significant (using the binomial test), and the trend is contradicted altogether when using ribosomal footprinting data[227]. Adapted from table S1, Wells et al.[1]

[i] If correct, the explanation for this given by Lim et al. is interesting[258]. As a result of coupling transcription and translation, genes at the 5' end of an operon will be transcribed and available to ribosomes before those at the 3' end. Since ribosome binding occurs almost as soon as transcription is underway, 5' genes will be actively translated for slightly longer than those at the distal end of the operon. However, to explain the magnitude of the effect they saw, they inferred that the efficiency of translation must be ~sixfold greater during transcription than after. A satisfying mechanistic explanation of this was not given at the time, but the recent discovery of the expressome complex[254] suggests a plausible answer.

2.2.4 Gene order matters most for lowly expressed protein complexes

Despite the majority of adjacent gene pairs having gene orders that correspond with assembly order, a significant proportion do not (~25% of pairs, disregarding whether or not gene order is conserved). What might account for the lack of correspondence in these cases? One possibility is that other constraints on gene order override protein complex assembly in these cases. Other studies have noted an analogous trend for gene order to match the order of metabolic pathways[250,259], so it is possible that similar biological phenomena could be influencing our results. However, analysis of gene ontology (GO) terms[260] did not reveal anything promising (figure A.6).

A more plausible explanation stems from our earlier observation that encoding subunits within operons is more common for lowly expressed protein complexes. In the immediate local environment of the active polyribosome, the concentration of interacting subunits will be high, but will drop off rapidly as the protein diffuses away from the site of translation. For proteins that are expressed at high levels, precise control of gene order might therefore be less important, since subunits still have a good chance of finding each other away from the site of translation. In figure 2.5, we show that in *E. coli* this does indeed seem to be the case, indicating that the minority of gene pairs in which gene order does not correspond to assembly order can mostly be explained by their high abundance. As for comparisons of operon-encoded and non-encoded complexes (figures 2.1, A.1), we see the same trend when considering all organisms in our dataset, and also when using absolute protein synthesis rates obtained from ribosomal profiling data (figure A.7).

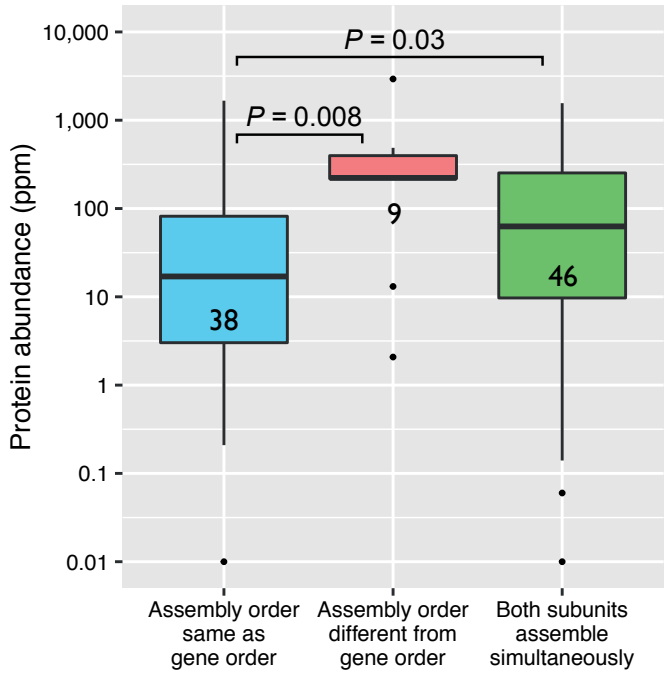


Figure 2.5. Gene pairs whose assembly order does not match gene order are highly expressed P-values were calculated with Wilcoxon rank-sum tests. Adapted from figure 4, Wells et al.[1] See also figure A.7.

2.3 DISCUSSION

These findings have a number of important consequences. From a technical standpoint, they demonstrate that computational predictions of assembly order are largely accurate and biologically meaningful. Together with earlier studies on the evolutionary conservation of assembly pathways[103,104], it is now fairly clear that in vitro pathways are similar to those in vivo - at the very least in bacteria. In eukaryotes it is known that chaperones play an extensive role in the assembly of complexes[261], and proteins that rely on these extensively probably deviate from easily predictable assembly pathways. In archaea on the other hand, we simply don't know enough about them to make confident assertions about the regulatory mechanisms they use to aid assembly of complexes. For one thing, we have substantially less structural and sequence information available for them[204], and for another, it is known that the structure of their operons differs from the canonical bacterial template[262].

An important assumption in this work is that expression control is primarily at the level of translation, with the mRNA-level expression of protein complex subunits being relatively constant within operons. In *E. coli*, which accounts for most of the data in this study, a substantial number of operons can produce alternative transcripts, thanks to the presence of internal promoters and terminators within operons[263]. In this work we have used the DOOR 2.0 database of prokaryotic operons, which distinguishes between operons and transcriptional units, i.e. a single mRNA output from an operon, and have used the latter in all analyses. Nonetheless, it is possible that some of the subtleties arising from overlapping transcriptional units within operons have been missed in this work.

Bearing these caveats in mind, we can nonetheless infer some interesting biological details about the assembly process. For example, the fact that there should be such a strong correspondence between gene order and assembly order for lowly expressed complexes implies that assembly must be taking place very close to the site of translation. This is because once proteins begin to diffuse away from the polyribosome, any knowledge about the order in which they were translated is rapidly lost. In the case of less abundant protein complexes, the effect of diffusion away from each other would be particularly strong, and as a result the selective benefit they get from minimising the spatial and temporal distance between interacting subunits is probably larger than for more abundant proteins.

Taking this argument further, it seems likely that in many cases operon-encoded complexes assemble co-translationally. In the specific case of the *Vibrio harveyi* heterodimeric luciferase complex, this has been demonstrated experimentally[44], and there is considerable evidence pointing to this being a common occurrence on a wider scale, not only in archaea and bacteria but also eukaryotes[230,264,265].

These results demonstrate that the pressure to assemble protein complexes quickly and efficiently has been a major constraint on the evolution of gene order in bacteria. However, eukaryotes are enormously different from bacteria and archaea in ways that prevent such a system as the one described here from being possible. Most obviously, the majority of eukaryotic species do not possess operons, and due to the existence of the nucleus, none couple transcription and translation. Indeed, the existence of operons appears to be largely decided by genome size and complexity, which

on average is much greater in eukaryotes than prokaryotes[266]. And yet, given the increased size of most eukaryotic cells, along with the diversity of intracellular organelles and compartments, it seems that the need for regulatory systems that facilitate protein complex assembly should be at least equal to that of prokaryotes, if not greater. The following chapters discuss a biological phenomenon that appears to be driven by such requirements.

3 DEGRADATION KINETICS OF PROTEINS ARE EXPLAINED BY ASSEMBLY OF PROTEIN COMPLEXES

3.1 INTRODUCTION

In the previous chapter, we saw how gene order in bacteria is optimised for assembly of protein complexes, in line with studies demonstrating that bacterial protein complex subunits are produced in proportions that correspond closely to the stoichiometry of the complex[227,267]. However, it is less clear that this is the case in eukaryotes, and the challenge they face in assembling protein complexes is compounded by the issues arising from their increased cell size and organisational complexity. Furthermore, as described in an influential paper by Papp, Pál and Hurst[46], there is a significant fitness cost to stoichiometric imbalances in the expression of eukaryotic protein complex subunits - this is known as the balance hypothesis. The effects of this dosage sensitivity manifest in phenomena such as the rapid degradation of excess ribosomal subunits[219,268], reduced noise in expression of dosage sensitive genes[269] and an under-representation of heteromer subunits in regions with high copy number variation[270] (CNV).

How does the eukaryotic cell balance the need for rapid and efficient protein complex assembly with the pressure to avoid the deleterious consequences of imbalanced expression of subunits? There are several ways to consider this question, but one that I have found to be important comes from a study on protein degradation kinetics, carried out in collaboration with McShane et al[2]. Here, I present the key results arising from this work, extended with some additional analyses relating to protein complex assembly and the balance hypothesis[46].

This project stems from a question posed by Matthias Selbach's lab, specifically: does the probability of a protein molecule being degraded remain constant over its lifetime? Early studies reached the conclusion that most intracellular protein degradation follows first order kinetics, implying that proteins have a fixed probability of being degraded at any moment in time[271,272]. However, it is not hard to imagine scenarios in which this assumption would be violated - for example, if proteins become unstable as they accumulate damage over time, or if they accumulate modifications marking them for destruction; alternatively, nascent proteins might initially be less stable than mature ones. There is some evidence for the latter scenario from early experiments showing that many proteins (including basigin and cystic fibrosis transmembrane conductance regulator) are degraded within the first few hours following synthesis[273–275]. Similarly it has been shown that ubiquitination of nascent proteins is common, and even occurs co-translationally with some regularity[276,277].

Considerable effort has already been invested in studying protein turnover using ribosome footprint profiling and mass spectrometric methods[224,225,278–280], particularly since the discordance between mRNA and protein abundance first became apparent[281,282]. However, most of the work thus far has focused on synthesis, since degradation is challenging to measure due to the difficulties in tracking newly synthesised proteins over extended time periods. To overcome this issue, my collaborators on the project performed pulse-chase experiments using the artificial amino acid azidohomoalanine (AHA)[283,284]. Combining this with SILAC mass spectrometry[136], they were able to track protein abundance changes in mammalian cells across seven time points spanning a 32-hour period. The results from this experiment yielded exciting insights into the nature of protein complex assembly in eukaryotes, lending themselves to a simple model that successfully predicts protein behaviour in aneuploid cells.

3.2 RESULTS

3.2.1 Measuring protein degradation kinetics

Early attempts at using pulse-chase experiments to follow protein degradation were hampered by the long pulse times required to label a sufficient proportion of proteins, which often considerably overlapped with the lifetimes of the proteins being studied[285]. This issue can be overcome with AHA - a bioorthogonal amino acid that has previously been used with comparatively short labelling times to track rapid changes in the proteome[286]. In order to measure changes in labelled protein abundance over time (figure 3.1A), mouse fibroblasts (NIH 3T3) were grown in heavy, medium and light SILAC medium, then pulse-labelled with AHA for one hour. The samples were then cold chased in AHA-free medium for durations specific to each of the three SILAC labels. Heavy cells were always harvested immediately after the pulse for use as the t_0 reference point for subsequent quantification. Medium and light samples were collected after cold chases of 1, 2, 4, 8, 16 and 32hrs, split across three independent replicates.

Samples were combined after harvesting to avoid introducing further batch effects and AHA proteins were enriched from this mixture on an alkyne agarose resin. The changes in protein abundance at each time point relative to t_0 were then quantified using LC-MS/MS. Mass spectra (MS1) for peptides from Filamin alpha (Flna), Cathespin L1 (Ctsl1) and Basigin (Bsg) are shown in figure 3.1B. Each of these three proteins displays a different decay profile, with Flna being very stable and decaying exponentially. Ctsl1 also decays exponentially but at a much faster rate, and ceases to be detected at time points beyond 8hrs. Finally, and most interestingly, Bsg appears to decay non-exponentially, with a rapid drop in abundance over the 8hrs hours, but very little change for the last 24hrs of the experiment. This is consistent with the earlier reports of Bsg in human cell lines undergoing rapid degradation during the first few hours of its life[274].

Controls to were carried out to verify that AHA did not induce premature translation termination or affect protein stability. Decay profiles were normalised using a set of abundant and very stable proteins to control for differences in cell number across samples. In total 5,257 proteins were quantified (figure 3.1), and after quality control (e.g. removing proteins with fewer than four data points) this was reduced to 3,605 profiles suitable for further analysis. Details on these aspects of

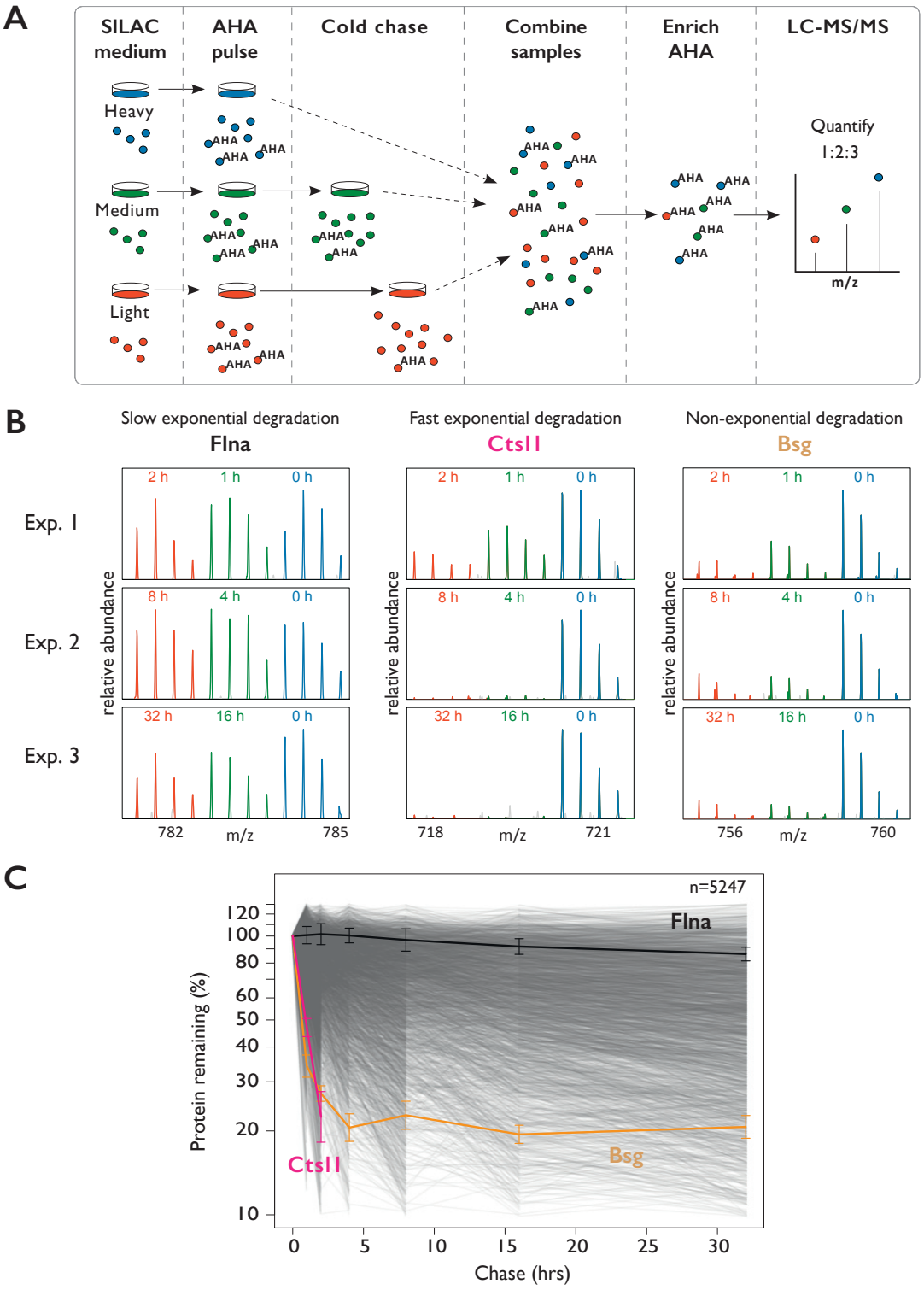


Figure 3.1.: Quantification of protein degradation kinetics by metabolic pulse-chase labelling
(A) Experimental setup for AHA pulse chase experiment. After a 1hr AHA pulse, each SILAC culture is harvested at different time point, with heavy cells being used as the t_0 reference point. After harvesting the cell cultures are combined, AHA proteins are enriched and then quantification of abundances carried out using tandem mass spectrometry. (B) This process was repeated in triplicate to obtain seven time points at intervals from 0 to 32 hours. Plots show mass spectra for three peptides from Flna, Ctsl1 and Bsg respectively. Relative changes in abundance are calculated from the area under different peaks. (C) Processed decay profiles for 5,247 proteins, with Flna, Ctsl1 and Bsg illustrating different possible degradation profiles. Adapted from figure 1, McShane et al.[2]

the work can be found in the methods and supplementary material of the original research paper[2].

3.2.2 Many proteins are degraded non-exponentially

Having acquired a large set of decay profiles, we next attempted to classify proteins according to their degradation profiles. Adapting a method previously used to study mRNA degradation[287], one- and two-state Markov models were fitted to the data for each protein (figure 3.2A). Under the first model, there is a single transition rate (and probability) K_A that describes the probability of a protein being degraded at any given moment in time. Under the second, there are assumed to be two states with different degradation rates, K_A and K_B. A protein in state A can thus either be degraded or transition to state B, at which point its degradation probability will change. The one-state model therefore describes exponential degradation (ED), whilst the two-state model describes non-exponential degradation (NED).

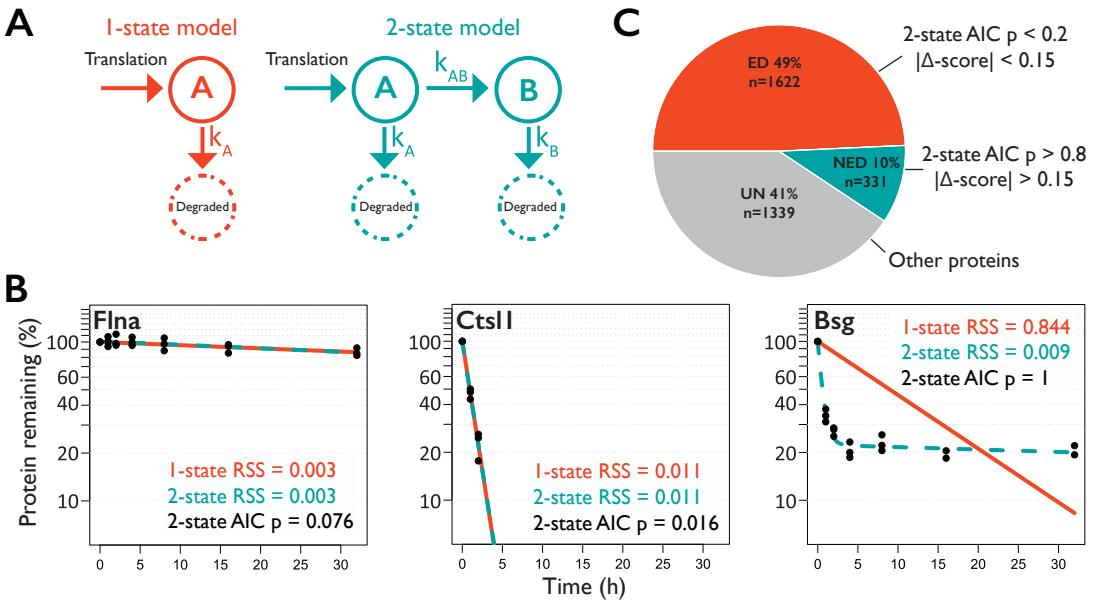


Figure 3.2.: Non-exponentially degraded proteins are common
(A) One- and two-state Markov models used to model protein degradation profiles. K_X describe transition probabilities between different states. (B) Markov models fitted to profiles of Flna, Ctsl1 and Bsg. Both Flna and Ctsl1 are equally well fitted by either model (i.e. $K_A \approx K_B$), as measured by the residual sum of squares (RSS), and therefore the simpler one-state model is selected. In contrast, Bsg is better fit by a two-state model, indicated by the favourable RSS and high AIC. (C) Under conservative thresholds for AIC and Δ-score 10% of proteins are degraded non-exponentially. Adapted from figure 2, McShane et al.[2]

To determine which was the better fit for each protein, we used the Akaike information criterion[288] (AIC), which imposes a penalty for additional parameters and therefore gives preference to the one-state model if both explain the data similarly well. However, AIC describes goodness of fit, but not the extent to which non-exponential degradation occurs; for example, a protein may be better fit by the two state model but only deviate weakly from a single degradation rate. To include this feature in classifications, a measure was developed (Δ-score) to assess the degree to which a pre-decided time point deviates from that expected under an exponential distribution. Specifically,

the distance of the data point at t_4 from the line fit by the one-state model between t_0 and t_8 (see methods in the original paper for details - appendix A.2). ED proteins were then classified as those which had a 2-state $AIC < 0.2$ and $|\Delta| < 0.15$, NED proteins as those with $AIC > 0.8$ and $|\Delta| > 0.15$, with proteins not meeting either criteria being undefined. Using the example cases of Flna, Ctsl1 and Bsg (figure 3.2B), we found that the first two are best fit by the one-state model, whereas the two-state model is preferred for Bsg, as one would expect from a visual assessment of the plots.

When we applied this classification scheme to the set of 3,292 proteins whose decay profiles passed quality control and for which Δ-scores could be calculated, we found that the majority of proteins (49%) were degraded exponentially, consistent with traditional models (figure 3.2C). Strikingly however, some 10% of proteins are best fit by a non-exponential model of degradation, despite the conservative nature of our classifier. A particularly unexpected finding was that in all NED cases, K_A was greater than K_B, indicating that these proteins are always initially rapidly degraded, before becoming more stable later in their lives. This suggests that, at least in the 32 hour time period we used, age- or damage-related destabilisation is rare, if it occurs at all.

3.2.3 NED proteins are degraded via the ubiquitin-proteasome system

There are two cellular systems concerned with degradation of proteins - the ubiquitin-proteasome system and the lysosomal system. To determine which of these was primarily responsible for degradation in the case of NED proteins, my collaborators inhibited each in turn using MG132, or wortmannin in combination with bafilomycin A1, respectively. Dimethyl sulfoxide was used a as a carrier control, and deviation from behaviour versus this control was measured by the change in Δ-score. Of these two treatments, MG132 lead to a marked decrease in NED character, whilst wortmannin in combination with bafilomycin A1 had no appreciable effect (figure 3.3). This indicates that initial rapid degradation of NED proteins is due to the ubiquitin-proteasome pathway, whilst the lysosomal system plays a negligible role.

3.2.4 NED proteins are enriched in heteromeric protein complexes

Based on analysis of a previously published set of manually curated protein complexes[289], my collaborators had found that NED proteins appeared to be over-represented in protein complexes. Building on this finding, I mapped proteins from the degradation dataset to protein structures from the PDB, revealing that approximately 70% of NED proteins are members of heteromeric protein complexes[i], which is a significant enrichment compared to either monomers or homomers (figure 3.4A). To exclude the possibility that this trend was driven exclusively by the ribosome (subunits of which are prevalent in our structural dataset), I repeated this analysis with ribosomes filtered from the dataset and obtained the same trend, with comparable statistical significance (figure A.8A).

[i] Though this may be a conservative estimate since protein complexes, particularly larger ones, are still somewhat undersampled in the PDB relative to monomeric or homomeric structures. Intriguingly, the number of novel homomeric and monomeric proteins released per year appears to be slowing down[290,291]. However, this is probably a result of increased interest in solving heteromeric structures caused by the technological developments discussed in chapter 1, rather than a sign of saturation of monomeric structure space.

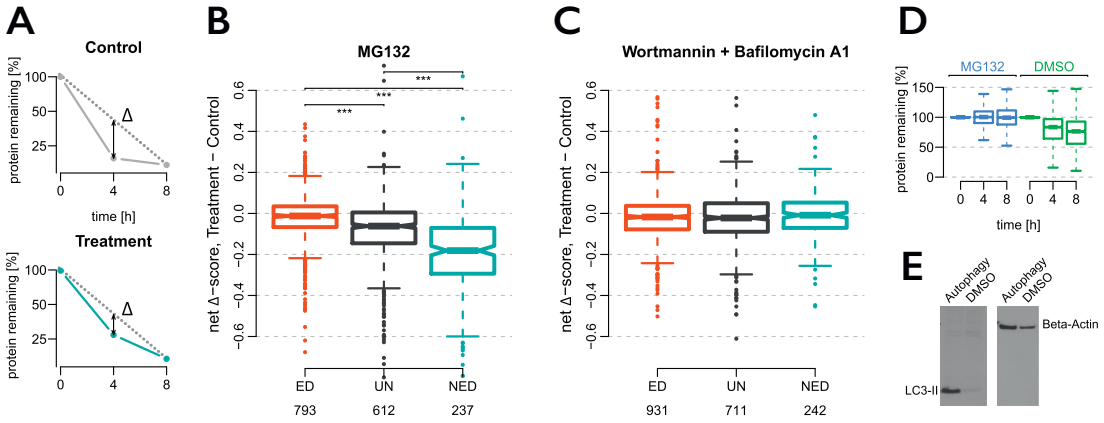


Figure 3.3.: NED proteins are degraded via the ubiquitin-proteasome system
(A) Example case showing hypothetical decrease in Δ-score caused by treatment at 4 and 8hr time points. (B-C) Results from treatment with MG132 or wortmannin in combination with bafilomycin A1, demonstrating significant reductions in NED protein Δ-score caused by the former. P-values calculated with one-sided Wilcoxon rank-sum tests; ***$p < 0.0001$. (D) Effect of MG132 on degradation of all proteins versus dimethyl sulfoxide control. (E) Control demonstrating inhibition of lysosomal degradation with the wortmannin-bafilomycin A1 combination (labelled 'Autophagy'). The accumulation of the autophagy marker LC3-II indicates that lysosomal degradation has been successfully blocked. Adapted from figure 4, McShane et al.[2]

Several independent studies have suggested that assembly of protein complexes stabilises the proteins involved[292–294]. These earlier studies are therefore supported by our findings that, firstly, NED proteins in our dataset exclusively transition to more stable states with age, and secondly, are enriched in heteromeric protein complexes.

Further supporting these studies, I observed a highly significant tendency for NED proteins to be found in large heteromers (measured in terms of number of unique subunits) compared to ED proteins (figure 3.4B). This is a reasonable finding if one assumes that as the size of complexes increases, so too will the proportions of subunits that are protected from degradation. Again, I repeated this analysis excluding ribosomes, and the enrichment of NED proteins in large complexes remained largely unchanged (figure A.8B).

Within individual protein complexes, NED proteins also tend to form larger interfaces than ED. Importantly, this trend holds when controlling for number of unique subunits (figure 3.4C), since larger complexes typically contain larger interfaces and from the last analysis we know that NED proteins are also enriched in these complexes. Intriguingly, despite the fact that NED proteins are not enriched in homomers (complexes with one unique subunit), there is nonetheless a significant tendency for homomeric NED proteins to form larger interfaces than ED.

Since there is a strong relationship between interface size and protein complex assembly order[104], it seemed likely that NED proteins would assemble earlier. Assembly order predictions were generated for the complexes in our structural dataset, and then normalised between 0 and 1, with lower numbers indicating earlier assembly. Since the assembly order of small complexes is relatively uninformative, I restricted the analysis to protein complexes with more than five subunits. Comparing the assembly position of NED and ED subunits, there is a weak tendency for the latter to assemble later, with undefined subunits somewhere between the two (figure 3.4D).

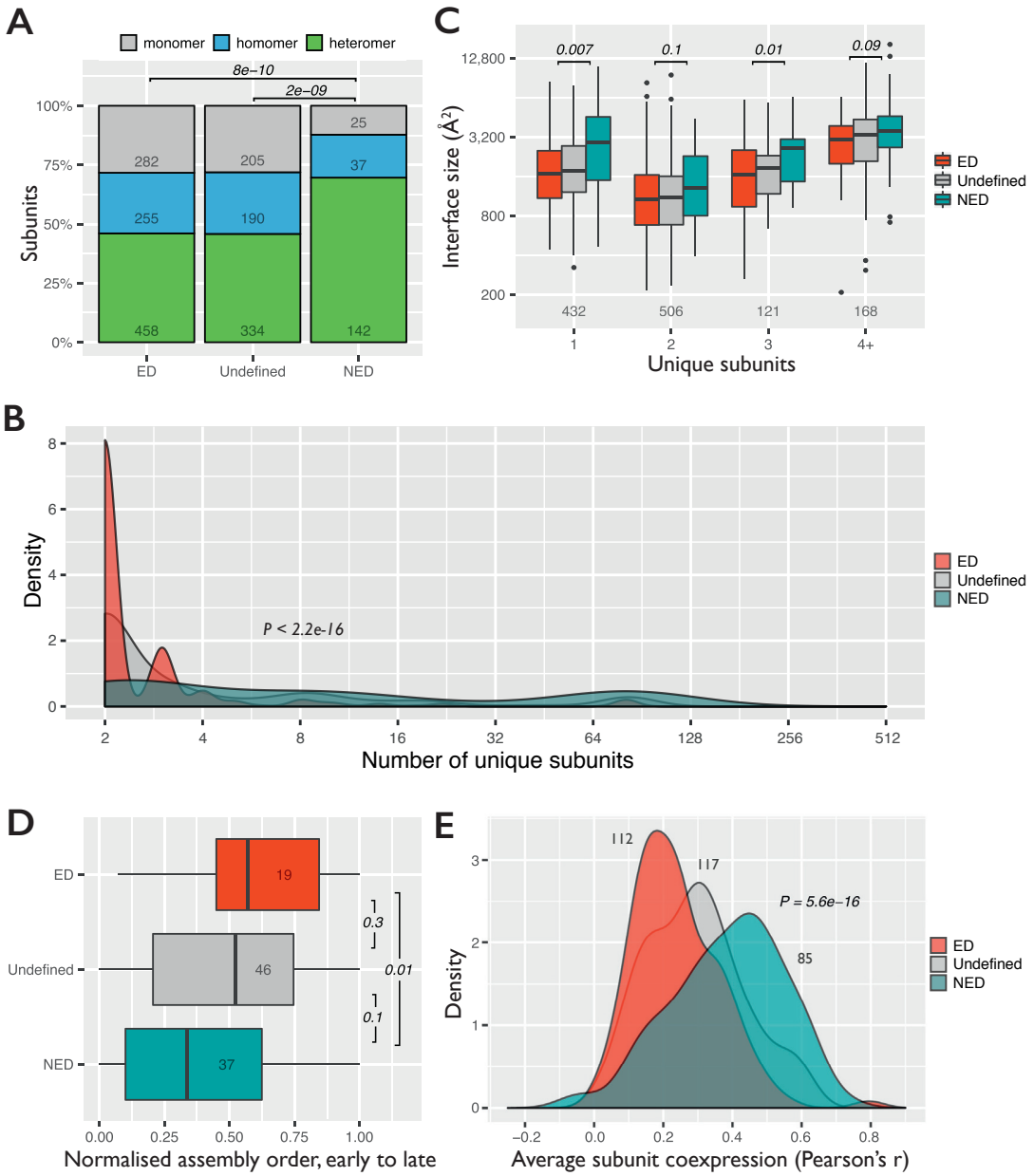


Figure 3.4.: NED proteins are enriched in heteromeric protein complexes
(A) NED proteins are proportionally much more common in heteromers than either monomers or homomers. P-values were calculated with Fisher's exact test, comparing the number of heteromeric subunits to monomeric or homomeric. (B) Considering only heteromeric protein complexes, NED proteins are more evenly distributed across the range of sizes, as measured by the number of unique subunits in each complex. P-value calculated using a modified Kolmogorov-Smirnov test to account for the discrete distribution of subunit counts - see R package 'dgof'[295]. (C) NED proteins tend to form larger interfaces, including when controlling for protein complex size. (D) NED proteins also tend to assemble earlier. Assembly order is predicted using the method described in Marsh et al.[104], and normalised between zero (early) and one (late). (E) Mean coexpression scores were calculated for each protein, based on all pairwise correlations with other proteins in the complex of interest. Expression of NED proteins are significantly more likely to be closely correlated with other subunits, compared against ED. P-values in panels C-E calculated with Wilcoxon rank-sum tests. Adapted from figure 5, McShane et al.[2]

Finally, I looked at the tendency for NED and ED proteins to be coexpressed with other members of the protein complex. To assign a single coexpression score to each protein complex subunit, I calculated the mean pairwise mRNA coexpression score between that subunit and all others in the complex. Comparing the two degradation classes, the expression of NED proteins is more tightly coupled to the rest of the complex (figure 3.4E).

3.2.5 Results are replicable across species and protein complex datasets

In order to ensure that these results were not specific to the mouse cell line being used (NIH 3T3), which has previously been shown to have an abnormal karyotype[296], my collaborators repeated the AHA-SILAC experiments using a human cell line (RPE1). This cell line was shown to have a normal karyotype, with the exception of partial trisomies of chromosomes ten and twelve. Their analyses of this data revealed that most proteins retain ED or NED character in human/mouse orthologues, and that Δ-score profiles of human versus mouse proteins are correlated (Pearson's correlation coefficient, $r = 0.39, p < 2.2x10^{-16}$). I then repeated the structural analyses described in figure 3.4 using the human dataset; this produced highly similar results (figure A.9), albeit with weaker statistical significance - probably as a result of the lower quality of the human data, with 49% of proteins passing quality control in this dataset compared to >60% in the mouse data.

Using non-structural data from CORUM[215] I repeated the coexpression analysis using mouse (figure 3.5A) and human data (figure A.10) datasets. In both cases there was a significant tendency for NED proteins to be more highly coexpressed. Since proteins from larger complexes such as the ribosome or proteasome tend to be highly correlated in their expression, I also split proteins into bins based on the number of unique subunits in the complex they were associated with (figure 3.5B). This showed that whilst there is a (not unexpected) tendency for NED coexpression to be stronger in larger complexes, overall the trend is not driven exclusively my complex size.

A final interesting analysis comes from looking at a mass-spectrometric dataset in which protein complexes were identified and quantified according to their stoichiometries and absolute abundances[106]. This approach allowed them to distinguish between 'core' protein complex subunits, which were identified by their closely matched stoichiometries both within complexes and across the entire cell. If NED proteins are indeed core subunits, as our analyses seem to suggest, then we would expect them to be enriched in the core set from the mass-spectrometric dataset. This turns out to be the case, with human NED proteins being significantly more likely to be identified as core interactors compared to ED (Fisher's exact test, odds ratio = 3.0, $p < 2.2x10^{-16}$).

3.2.6 Protein complex assembly explains degradation kinetics

Collectively, these observations suggest that NED proteins tend to be core subunits within protein complexes, whereas ED proteins are more likely to be monomeric/homomeric or participate as peripheral subunits within heteromers. A simple model that would explain why core members of subunits would display non-exponential degradation kinetics is shown in figure 3.6A. If core subunits are more abundant than peripheral subunits, then there would always a fraction of proteins unable to assemble into complexes, and thus subject to rapid degradation. My collaborators

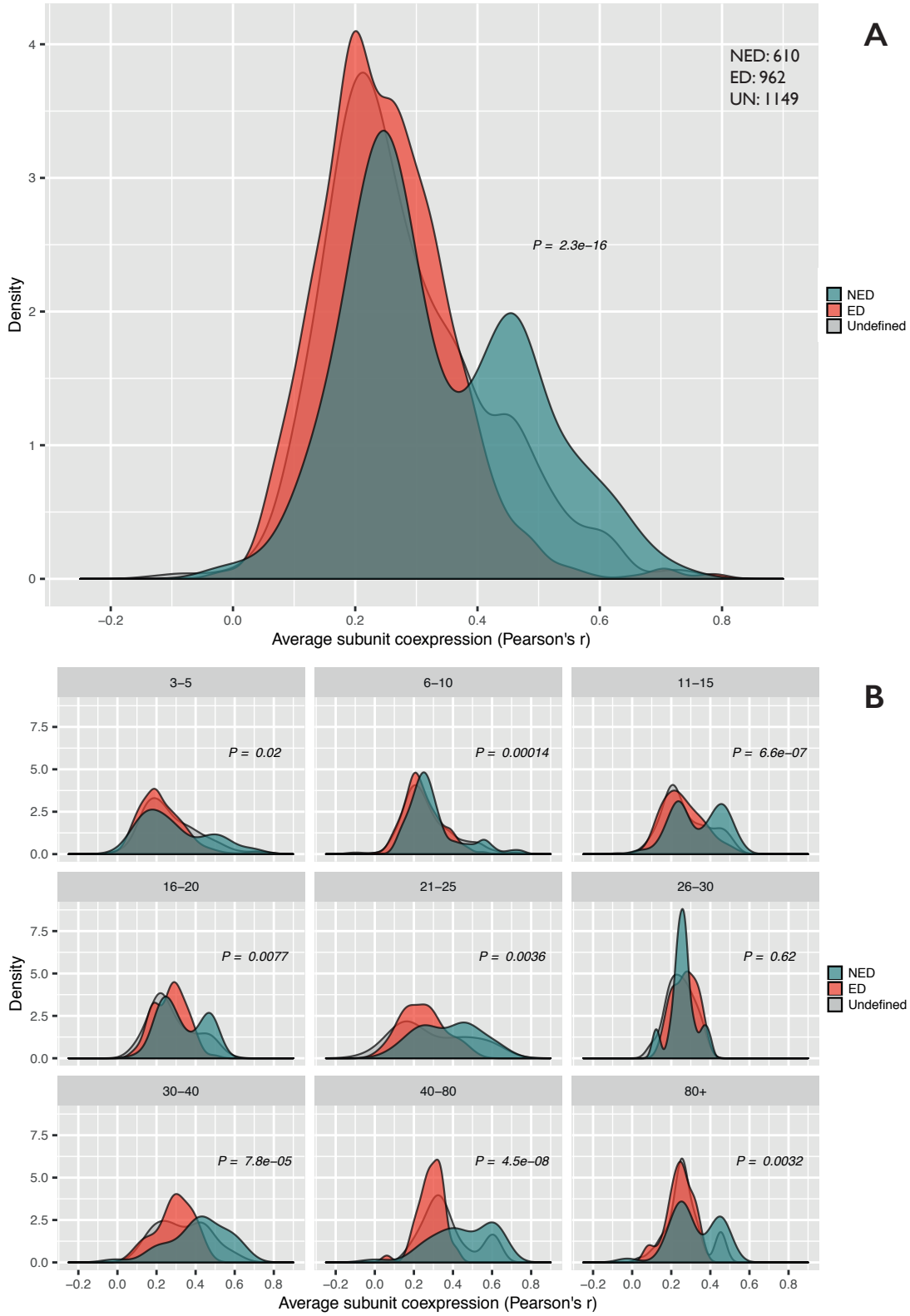


Figure 3.5.: Increased NED protein coexpression is not unique to structural data
(A) Replicate of figure 3.4E using mouse data and experimentally validated protein complexes from CORUM[215]. The bimodal distribution of NED subunits is likely due to the strong influence of the ribosome, which is tightly regulated and contributes a substantial number of all proteins in the dataset. (B) This trend holds when controlling for the size of protein complexes, measured in terms of number of unique subunits. P-values calculated with Wilcoxon rank-sum tests.

estimated absolute protein abundances after the AHA pulse using intensity-based absolute quantification[224] (iBAQ), and mapped these to the set of protein complexes they used previously[289]. Normalising abundances relative to the mean of each complex, we found that NED proteins within complexes were significantly more abundant than ED (figure 3.6B), consistent with our model. In addition, the second-state degradation rates of NED proteins were more similar to those of ED within complexes.

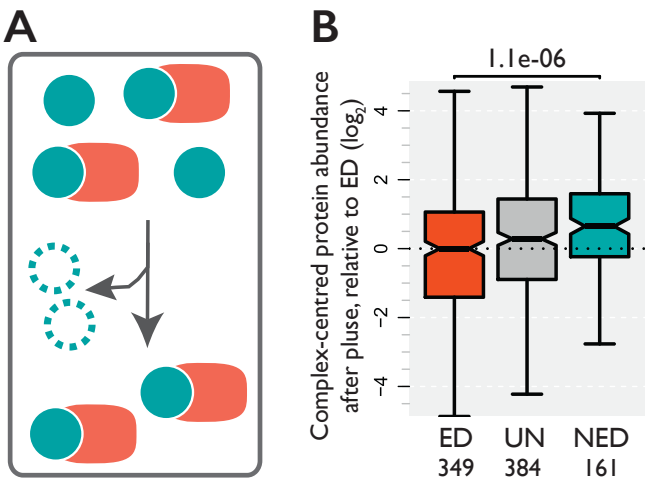


Figure 3.6. NED proteins are produced in excess in heteromeric complexes
(A) Model showing how overproduction and subsequent degradation of excess NED subunits facilitates protein complex assembly. (B) NED subunits are significantly more abundant per complex than ED proteins. Log2 abundance fold change of each subunits calculated against the mean abundance of each complex. P-value calculated with Wilcoxon rank-sum test. Adapted from figure 5, McShane et al.[2]

The finding that NED proteins (core protein complex subunits) tend to be over-produced relative to ED (more peripheral) is somewhat paradoxical in light of the dosage balance hypothesis described by Papp et al[46]. This hypothesis states that changes in gene expression that affect the stoichiometric imbalance should be deleterious, and indeed there is much evidence to support this; however much of it relates to large changes in gene expression caused by copy number variants. If excess heteromeric subunits are rapidly degraded, as demonstrated by the results presented in this chapter, then this suggests that smaller fluctuations in expression might be better tolerated by heteromers than monomeric or homomeric proteins - more specifically, increases in expression should be better tolerated than decreases, since these can be degraded. To test this idea, I mapped expression quantitative trait loci (eQTL) data from the Genotype-Tissue Expression[297] (GTEx) project onto a set of human protein complexes; if correct, then heteromers should on average have more eQTLs per gene. Separating these out by quaternary structure category revealed that, contrary to this hypothesis, the number of significant eQTLs per gene is lower for heteromeric subunits, in accordance with results supporting the balance hypothesis (figure 3.7). Moreover, there are not meaningful differences between the number of upregulating and downregulating eQTLs across different quaternary structure types.

3.3 DISCUSSION

This study of protein degradation kinetics revealed that many proteins are degraded non-exponentially, indicating that examples previously reported in the literature are in fact part of a widespread phenomenon. To explain these observations, we proposed a model in which protein complex assembly stabilises those proteins involved, with excess subunits being degraded. As we shall see in the

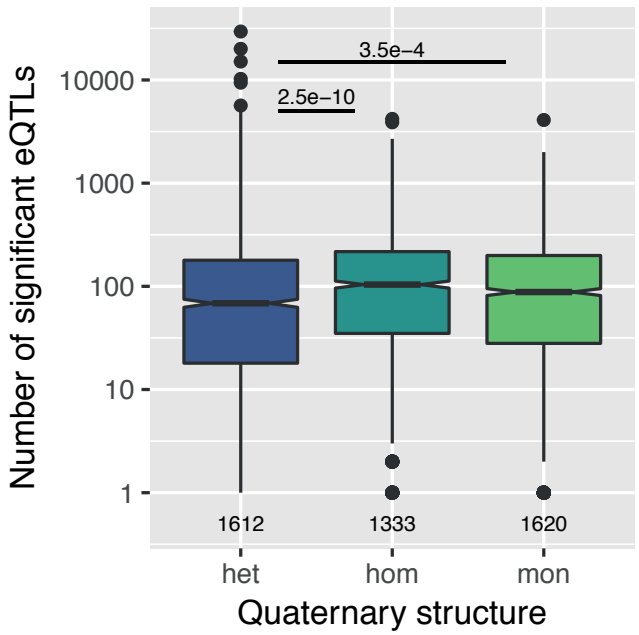


Figure 3.7. eQTLs are less frequent for heteromeric proteins
If rapid degradation of excess heteromer subunits acts to increase robustness in the expression of protein complexes, then we would expect eQTLs to be better tolerated in these proteins. However, this does not appear to be the case, as evidenced by the lower frequency of eQTLS per gene in heteromers. P-values calculated with Wilcoxon rank-sum tests.

next chapter, this model successfully explains protein attenuation in aneuploid cells. In addition, the findings support the idea that many heteromers, particularly larger ones, possess sets of core subunits, the structural and behavioural properties of which differ markedly from peripheral or non-essential subunits. However, the work also raises questions: for example, what are the mechanisms that explain why assembly into complexes stabilises proteins, and why are eukaryotic protein subunits not produced at levels corresponding to their stoichiometry, as appears to be the case in bacteria[227]?

There are a few mechanisms could explain why proteins are stabilised upon binding. The simplest is that there is a safety-in-numbers effect, whereby proteins in large complexes that bury more surface area are less accessible to the proteasome. This is certainly consistent with the tendency of NED proteins to have larger interfaces and assemble earlier. However, this alone does not explain why NED subunits are initially degraded faster than ED subunits within complexes. One factor that is probably important is ubiquitination - it would be interesting to see if there is enrichment for ubiquitination sites in NED proteins or protein complex subunits in general, as some evidence seems to point to[298]. If this is the case, are ubiquitination sites enriched or depleted in protein interfaces, and can ubiquitinated proteins still assemble into functional complexes?

The published work this chapter derives from discusses a number of possible reasons why NED proteins might be overexpressed relative to ED. One interesting idea is that overexpression of core subunits represents a simple mechanism for controlling levels of the holocomplex. If the central NED subunits are constitutively overexpressed as a single, correlated group relative to late-assembling ED subunits, then the levels of the fully assembled complex can be modulated by that of the ED subunits alone. This fits with the observation that expression of ED subunits is often decoupled from that of the rest of the complex. On the other hand, since most heteromeric subunits are NED, the finding that eQTLs are still less common in heteromers suggests that NED proteins

are not necessarily any more robust to changes in expression.

A simpler alternative is that overexpression of early-assembling core subunits is a necessity in ensuring efficient assembly, since the encounter of interacting proteins is concentration dependent, and eukaryotes do not benefit from the high local concentration of subunits produced by encoding heteromers in operons. This scenario goes some way to explaining why eukaryotic subunits are not expressed at stoichiometric ratios, and does not contradict the balance hypothesis, since proteins may be expressed non-stoichiometrically, but nonetheless at tightly controlled levels.

In summary, this study shows that non-exponential degradation of proteins is commonplace, and is driven by the ubiquitin-proteasome system and the inherent instability of core heteromeric subunits. It is important to note that whilst NED character is conserved across species, it is not an inherent property of the proteins themselves, but rather an emergent phenomena caused by the requirements of protein complex assembly in eukaryotes. This can be shown by the fact that NED character decreases in response to repressing the proteasome, and also by the way in which proteins behave in aneuploid cells.

4 AUTOSOMAL DOSAGE COMPENSATION IN ANEUPLOID CELLS

4.1 INTRODUCTION

Cancer cells are renowned for their high levels of chromosomal instability and unusual, sometimes bizarre karyotypes. The state in which a cell has an abnormal number of chromosomes is known as aneuploidy, and occurs to degrees in all known cancers. However, it is common even in non-cancer cells, and is a pervasive feature in most eukaryotic organisms. Isolates of wild yeast strains for example have been found to harbour a variety of different karyotypes,[299]. In humans, approximately 0.1% of the population carries an extra copy of chromosome 21[300], a karyotype that famously results in Down's syndrome. In most cases however, aneuploidies that have been acquired through the germ-line or early in development are lethal - highlighting this, a recent study of spontaneous miscarriages found that approximately 45% were the result of aneuploidies[301], and the true figure is probably higher still, since studies in mice suggest that mosaic aneuploid embryos fail to develop much beyond gastrulation, and thus would often pass clinically undetected[302].

Mechanistically, aneuploidy is the result of chromosomes failing to separate properly during cell division, either through non-disjunction of sister chromatids or delays in the movement of chromosomes to opposite poles of the cell during anaphase[303]. For a cell that gains a chromosome, the immediate effect is a doubling of the copy number of all the genes residing on it, thus leading to a significant increase in mRNA and protein production. However, a recurring feature noted in several studies is that a considerable number of these proteins are attenuated compared to their expected abundances.[304–306].

This attenuation is generally thought to be caused by post-translational degradation of excess protein complex subunits, since attenuated proteins are enriched in protein complexes, and their mRNA transcript abundances scale correctly with copy number. Strongly supporting this, the model presented in the previous chapter explaining non-exponential degradation of proteins also correctly predicts the attenuation of proteins in aneuploid cells. In this chapter, I will briefly elaborate on this idea and attempt to explain some of the similarities and differences between non-exponential degradation in wild type cells and protein attenuation in aneuploid cells.

4.2 RESULTS

4.2.1 Attenuation of protein complex subunits is unique to heteromers

First, to confirm that attenuation of protein complex subunits is a feature unique to heteromers, I made use of data from Dephoure et al.[305]. This dataset describes the relative fold change in the abundance of 2,581 genes from disomic *S. cerevisiae* strains. Using the same definition given by Dephoure et al., attenuated proteins are those satisfying:

$$\log_2\left(\frac{disomic}{wildtype}\right) \leq 0.6$$

where $disomic$ and $wildtype$ refer to the abundance of individual proteins in the two different cell populations. This value ensures that the observed SILAC ratio is at least three standard deviations away from the expected mean value of 1.0 (equivalent to doubling abundance). One of the key findings from this paper was that attenuated proteins were enriched in a set of multi-subunit complexes obtained from a dataset of yeast complexes compiled from the experimental literature[307]. After replicating this finding using the same dataset (figure A.11), I mapped this aneuploidy dataset onto structures from the PDB. This structural dataset confirmed these earlier observations that attenuation of expression is more common for members of multi-subunit protein complexes, and demonstrated that attenuation was restricted to heteromeric subunits (figure 4.1).

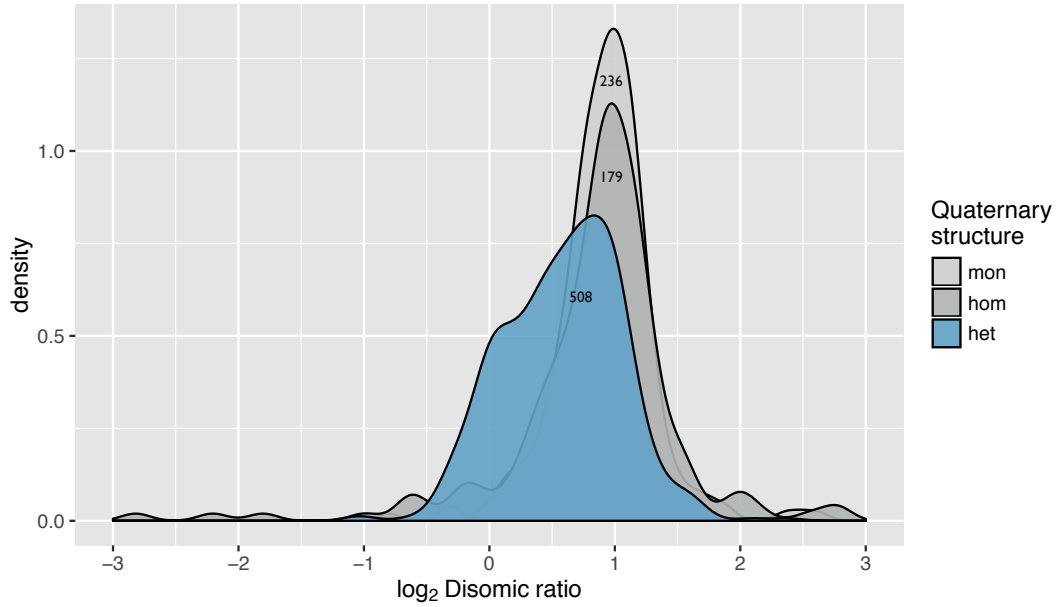


Figure 4.1.: Attenuation of protein complexes is unique to heteromers
Members of protein complexes are significantly more likely to be attenuated upon doubling of gene copy number. However, this effect is almost entirely driven by heteromeric complexes, with disomic ratios that differ significantly from that seen in monomers (median log2 disomic ratios of 0.623 and 0.948 respectively, Wilcoxon rank sum test p-value = 4.315e-22). In contrast, homomers (median 0.942) behave much the same as monomers and the difference between them is not at all significant.

Importantly, this observation is consistent with the model presented in chapter 4, in which non-

exponential degradation is a consequence of excess protein complex subunits being degraded more rapidly than bound ones. Adapted to this situation, disomic ratio is dependent on the ratio of bound to unbound subunits. Duplicating a single subunit in a heteromer increases the unbound, unstable fraction of that protein, which is promptly degraded, leading to a lower disomic ratio. Proteins that are predominantly monomeric should not experience any systematic attenuation, since their degradation rate will not be affected by binding partners and should remain roughly constant. Likewise, subunits of homomeric complexes will not be attenuated since changes in gene copy number will not cause stoichiometric imbalances.

4.2.2 Similarities and differences between wild type subunit degradation and aneuploid attenuation

Many of the features we see in ED and NED proteins reappear here, suggesting that they are different manifestations of the same underlying biological phenomena. For example, as complex size increases, the average disomic ratio of subunits decreases (figure 4.2), analogous to the enrichment of NED proteins in large complexes. As before, this is probably explained by the fact that larger proteins have a greater proportion of obligate subunits that are protected from degradation once bound.

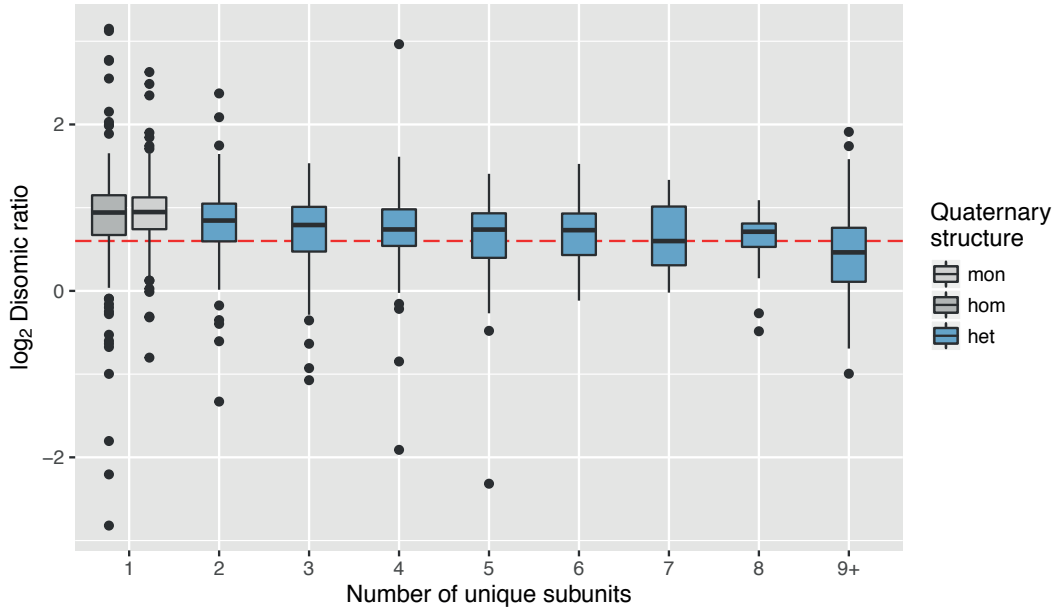


Figure 4.2.: Degree of attenuation increases with increasing complex size
The abundance of monomers and homomeric subunits is largely determined by gene copy number, with $\log_2$ disomic ratios being approximately normally distributed about 1. Heteromeric subunits however become increasingly likely to be attenuated as the increasing number of unique subunits. "Significant" attenuation is defined here by a threshold value of < 0.6, highlighted by the dashed red line. Structural and non-structural datasets have been combined here.

In contrast, if one imagines a typical heterodimer, the observed attenuation for each subunit will be determined by the degree to which its binding partner buffers that subunit. This in turn is dependent on binding affinities, dissociation rates and starting concentrations. For example, if the

starting concentration of A is much higher than B, then even doubling [B] will have little effect on the amount of unbound B, and thus it will appear to be non-attenuated.

One observation raised in Dephoure et al. is that almost all complexes have at least one non-attenuated subunit. An interesting idea mooted to explain this is the possibility that complexes require at least some stable subunits to act as a scaffold for the rest of the complex. If this were the case, then we would expect to see non-attenuated proteins being amongst the first to assemble. However, if we compare the disomic ratio of those subunits that assemble first against those that assemble last (figure 4.3), we find a clear (albeit weakly significant when controlling for number of unique subunits) tendency for early-assembling subunits to be attenuated. This finding matches the observations showing that NED subunits tend to assemble earlier than ED subunits, and is incompatible with the idea of stable proteins acting as scaffolds for protein complex assembly. Consistent with this, attenuated proteins also tend to form larger interfaces.

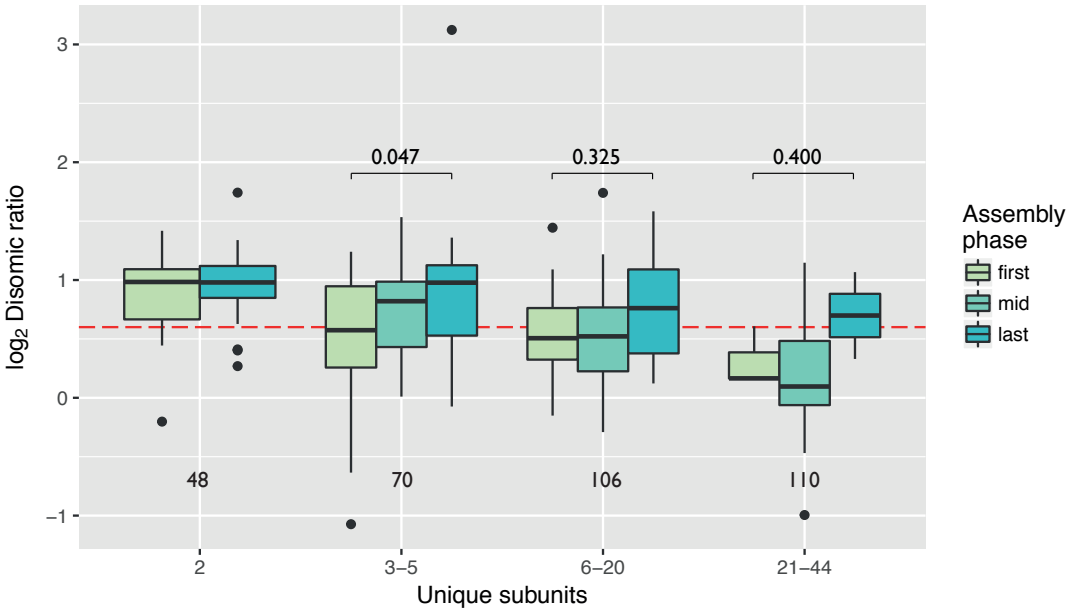


Figure 4.3.: Subunits that bind late to the complex are less likely to be attenuated
Red line indicates 'attenuated' threshold. P-values are Wilcoxon rank sum tests indicating the difference in disomic ratio between the first and last subunits to assemble. 'Mid' subunits are all those that are neither first nor last to assemble. The non-significant values are probably mostly due to the fact that the 'first' and 'last' subunit numbers are highly limited by the number of available PDB structures, which is relatively low for the larger complexes.

An interesting difference between NED proteins and attenuated proteins is in their abundance. At first glance, one might expect to see a similar case as for the NED proteins, where within complexes NED proteins tend to be more abundant than ED. However, when we look at this for the disomic ratio data there is no significant difference between proteins whose abundance is attenuated and vice versa (figure A.12). On reflection, this makes sense - in wild type cells, the more abundant protein will be degraded non-exponentially, whereas in aneuploid cells, the natural abundance of the protein is rendered insignificant compared to the effect of copy number changes.

Less easy to explain however is the fact that attenuated proteins appear to show a greater propen-

sity to disorder than non-attenuated (figure 4.4). This is in contrast to NED and ED, in which the latter were found to be more disordered[2]. One possibility that could explain this discrepancy is that in aneuploid cells - in which protein abundance has been drastically increased - disordered proteins are simply aggregating instead of being degraded. Though I have not yet tested this idea formally, there is evidence to suggest that disorder is correlated with aggregation propensity[308].

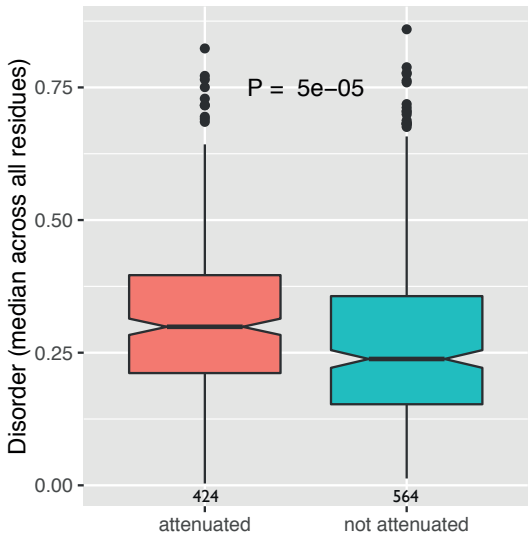


Figure 4.4. Attenuated proteins show increased disorder
A single disorder score for each protein in a set of protein complexes was produced by taking the median value from all residues, where per-residue disorder scores were predicted using IUPred[309]. Proteins were taken from a set of complexes produced by combining structural data with that from the Pu et al. dataset[307]. P-value calculated with Wilcoxon rank-sum test.

4.2.3 Aneuploidy leads to increased heteromeric protein aggregation

Protein aggregation is in an important aspect of aneuploidy that has not yet been explored in a high-throughput manner. Beyond the possibility that it could explain this discrepancy between disorder propensity in NED and attenuated proteins, it is also possible that the apparent attenuation of some proteins could be due instead to aggregation. However, measuring this is technically challenging due to the difficulty of isolating and quantifying the aggregated fraction of cells. Nonetheless, a pilot study carried out in collaboration with the Amon lab at MIT has produced some interesting results that merit further exploration in the near future.

The protocol followed for this pilot is essentially the same as that described in Dephoure et al., with the key difference being that instead of carrying out measurements on the whole cell lysate (in 8M urea), the lysate is spun down and the aggregate fraction extracted for further use. Disomic:wild type ratios were calculated in the same manner as before, with proteins being classified as either 'aggregating' or 'highly aggregating'. The threshold for the latter is set semi-arbitrarily at a disomic ratio ≥ 1.5, where a value of one would indicate a straightforward doubling of protein abundance in the aggregate fraction.

At least in aneuploid cells in which chromosome copy numbers have been increased, one would expect to see an increased propensity for aggregation across the cell, and particularly for proteins on affected chromosomes. This is certainly the case, and furthermore, there is an increased tendency for heteromeric proteins to aggregate more than expected upon copy number duplication (odds ratio = 2.392818, p-value = 0.038, comparing heteromeric to monomeric proteins). As was the case for attenuation in whole-cell measurements, there was no significant difference between monomeric

and homomeric proteins.

I then compared those proteins which were found to be highly aggregating with those that are attenuated and found a weak tendency for the different classes to be mutually exclusive (figure 4.5). This suggests that proteins are either aggregate extensively, are actively attenuated, or behave neutrally and simply double their abundance in either soluble or insoluble fractions of the cell. Importantly, this tells us that attenuation in aneuploidy is unlikely to be an experimental artefact caused by the aggregated fraction of proteins not being efficiently picked up in whole-cell lysates.

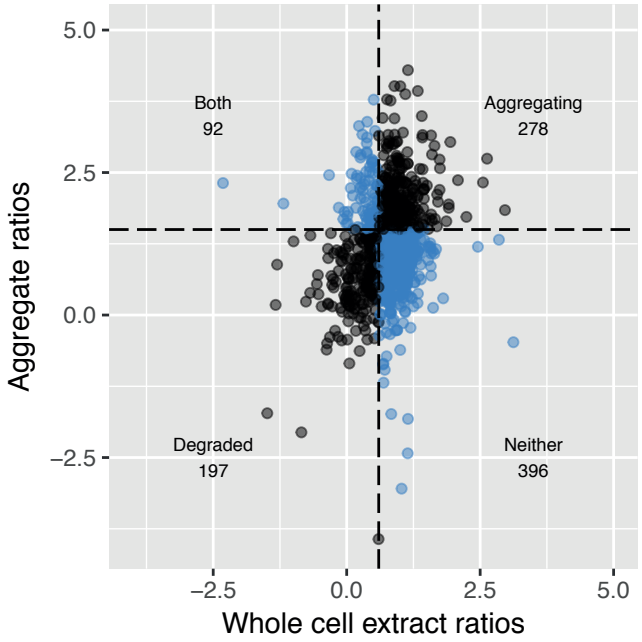


Figure 4.5. Comparison of attenuation and aggregation propensity
Separating proteins into groups according to whether they are attenuated or highly-aggregating demonstrates that for proteins to be both heavily attenuated and highly aggregated is unusual.

Intriguingly, there are no obvious features that suggest why some proteins are rapidly degraded whilst others are prone to excessive aggregation. Like attenuated proteins, highly-aggregating proteins share features with NED proteins and show indications of being core protein complex subunits, as indicated by attributes such large interfaces and being highly coexpressed within complexes (figure 4.6).

4.3 DISCUSSION

Supplementing the finding that NED predicts the response of proteins to aneuploidy, these analyses go further in explaining the connection between attenuation in aneuploid cells and non-exponential degradation in normal cells. Both sets of results suggest that the assembly of a complex is not centred on stable scaffold proteins, but rather on a core set of unstable, highly expressed subunits. The first subunits to assemble tend to be rapidly degraded, with significant stabilisation occurring upon binding, as indicated by the NED data and again here. In contrast, the last subunits to bind tend to be comparatively stable in both bound and unbound states. The take home point from this is that both NED and attenuation behaviour are caused by rapid decay of excess protein complex subunits - this is further supported by a recent study revealing highly consistent observations with CNV proteins from tumour sample[306].

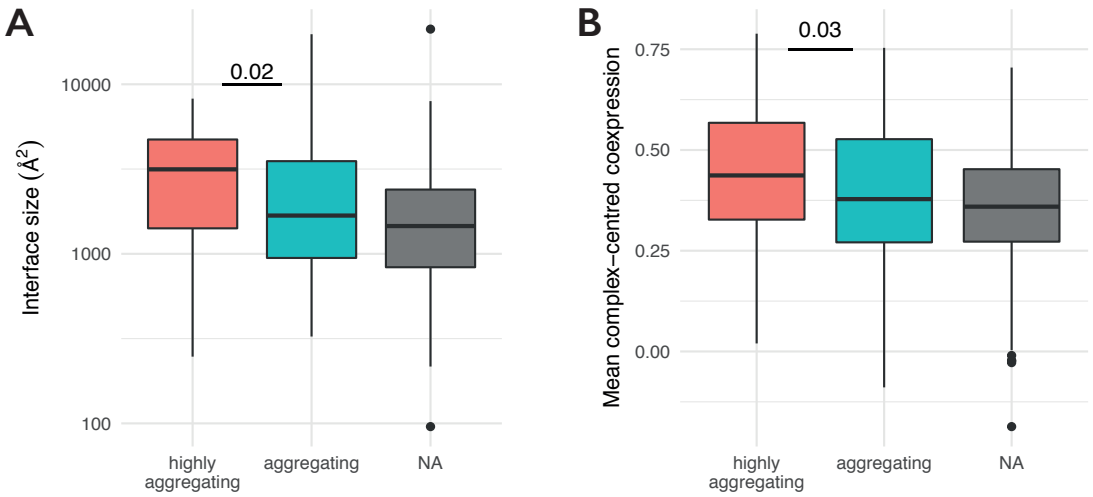


Figure 4.6.: Features of aggregating proteins
(A) Highly aggregation prone proteins form larger interfaces. (B) And are more highly coexpressed with other subunits within complexes. In both cases NA stands for not available, but this does not necessarily mean non-aggregating, since it could include proteins which only form aggregates in either wild type or aneuploid states. P-values calculated with Wilcoxon rank-sum tests.

The clearest difference between attenuation such as that seen in aneuploid cells or tumours, and non-exponential degradation common to all cells, is in the fact that NED is the result of relatively slight differences in protein expression within complexes, whereas attenuation is the result of much larger increases in expression affecting all of the proteins on affected chromosomes. However, there are still important questions to be tackled in understanding the behaviour of aneuploid cells - we have started to address one of these with a preliminary study of changes in aggregation propensity caused by aneuploidy.

One important point to note is that classifying proteins as 'aggregating' or 'highly aggregating' does not take into account proteins which simply don't aggregate at all, or proteins which do not aggregate in wild type cells but do in aneuploid. Similarly, whilst the aggregating fraction of proteins is certainly much smaller, with the data we currently have it is not possible to make quantitative comparisons of relative amounts of aggregated vs. non-aggregated proteins.

Additionally, there are work-in-progress issues in distinguishing between technical and biological noise in measuring the change in abundance of the aggregating fraction of proteins. As the quantity of protein that can be purified from the aggregating fraction of the cell is very small, there is substantial noise in the measurements. The data from each experiment is therefore normalised by mean-centring (see methods), such that the mean of data points from each experiment is identical.

This reduces the variance in measurement of aggregate disomic ratios (supplementary figure A.13), but comes at the expense of being able to investigate the relationship between chromosome size and aggregation propensity. For example, it is possible that duplicating larger chromosomes leads to a greater overall increase in aggregation propensity across the cell. Additional replicates of selected chromosomes are being planned to better establish the extent to which noise in the data is attributable to technical differences compared to biological ones.

The most pressing question that remains from this work is why some proteins are attenuated

whilst others aggregate. Work from other groups[220,310], has shown that members of protein complexes are aggregation-prone. Our data supports this, since we see that highly aggregating proteins are enriched in heteromeric complexes and many attributes of aggregated proteins are shared by NED and attenuated proteins. The simplest reason that explains why some proteins are aggregated therefore is seems to be simply that they are less efficiently degraded. Testing this hypothesis will be a priority as this study goes forward.

5 HAWK PROTEINS: A PARALOGOUS FAMILY OF EUKARYOTIC SMC-KLEISIN REGULATORS

5.1 INTRODUCTION

The evolution of protein complexes differs from that of individual proteins in that functional adaptations, beneficial or otherwise, can be driven by gain and loss of subunits, in addition to the mechanisms that operate on individual proteins[291,311,312]. In this chapter I will discuss a case in which the replacement of one family of subunits with another has played an essential role, namely in the evolution of condensin and cohesin.

The SMC-kleisins (Structural Maintenance of Chromosomes) are an ancient family of protein complexes found in archaea, bacteria and eukaryotes. As the name suggests, these complexes have roles relating to maintenance of chromosomal architecture and DNA integrity, and operate at various stages of the cell cycle[313,314]. Structurally, they are distinguished by their unusual tripartite ring formation, comprising two SMC arms that form a V-shaped dimer, linked by a largely disordered kleisin subunit (figure 5.1). In eukaryotes there exist three main subfamilies of the SMC-kleisins: condensin, cohesin, and Smc5/6, whereas prokaryotes have at least three that we know of - MukBEF, MksBEF, and Smc/ScpAB.

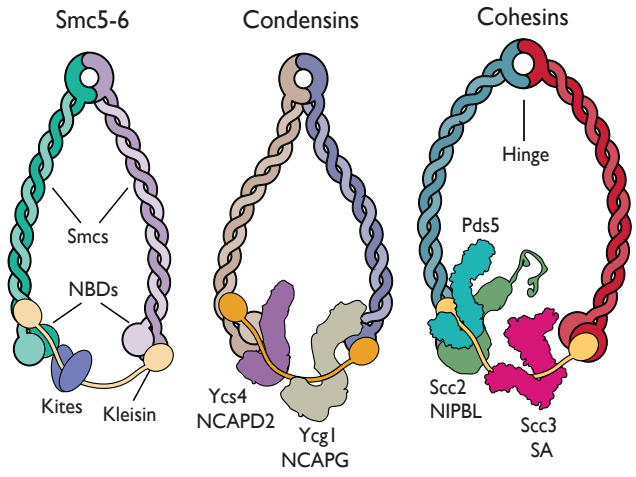


Figure 5.1. Eukaryotic members of the SMC-kleisin family of protein complexes
Eukaryotic SMC-kleisis are formed from SMC heterodimers linked at a 'hinge' domain. Each arm is connected via kleisin subunits (yellow) which bind to globular nucleotide binding domains (NBD). Various regulators interact with the kleisin - kites in the case of Smc5-6 and hawks in the case of condensin and cohesin. In cohesin, Pds5 and Scc2 are thought to compete for binding space on Scc1 (kleisin)[315]. Adapted from figure 1, Wells et al.[3]

Condensin and cohesin are involved in many cellular processes, including two hallmark features of eukaryotic cell division - condensation of chromosomes and sister chromatid cohesion[316]. Both complexes have been studied extensively, and condensin in particular is currently the subject of intense interest due to its central role in loop extrusion, a beautiful model that looks increasingly

likely to be the correct mechanistic explanation for chromosome condensation[317–320]. In contrast to condensin and cohesin, the structure and function of the Smc5/6 complex is less well understood, but is thought to be more closely related to prokaryotic SMC-kleisins[321] and is involved in DNA repair, specifically the resolution of recombination intermediates during mitosis and meiosis[322,323]. In practice however, across the extensive literature on the family (reviewed recently by Frank Uhlmann[324]) the evidence indicates that there is a considerable degree of functional overlap between different members of the family, in both prokaryotes and eukaryotes.

In addition to the SMC and kleisin subunits, numerous regulatory proteins are also associated with the complexes. One such group of regulators is the Kite proteins (Kleisin interacting winged-helix tandem elements), which form dimers that interact with the SMC-kleisin rings in bacteria, archaea, and the eukaryotic Smc5/6 complex[321]. However, these are missing from condensin and cohesin, which instead interact with a number of proteins containing tandem HEAT (Huntingtin, EF3, PP2A, TOR1) repeat motifs[325].

HEAT repeat proteins are a highly diverse family that is involved in cell processes ranging from signalling (beta-catenin in the Wnt pathway[326]) to intracellular transport (clathrin adaptors[327] and karyopherins[328]). Though it has been known for many years that a subset of these HEAT proteins are involved in the regulation of condensin and cohesin[329], neither their evolutionary history nor the extent of their coverage in the SMC family has been investigated. Building on the recent description of the Kite family, we asked whether this subset descends from a single ancestral HEAT repeat protein, and when this ancestor appeared.

Due to the nature of HEAT repeat proteins and repetitive sequences in general, these question presents non-trivial technical difficulties. Repetitive sequences can and often do diverge rapidly upon duplication[330]; illustrating this, the average sequence identity between mammalian HEAT repeat proteins and insect orthologues is just 13%[331]. This makes accurate sequence alignment challenging, and therefore classical methods for homology detection fail on all but the most similar of these proteins. To tackle this problem I developed a computational approach based on extensive profile HMM searches and network clustering. Using this method, we were able to answer important questions about the origin of the HEAT protein regulators of condensin and cohesin, and thus we propose naming the group Hawks, i.e. HEAT proteins associated with kleisins.

5.2 RESULTS

5.2.1 Resolving evolutionary relationships between repeat proteins

To detect potential paralogues of the candidate hawk proteins, I first used HHblits[332] to search the *Saccharomyces cerevisiae* proteome (UniProt ID UP000002311) for proteins with strong sequence similarity to the candidates. This immediately showed highly significant alignments between different members of the hawk group - alignments that were not detectable by less sensitive, but more widely used methods such as PSI-BLAST[333]. However, these searches also revealed significant similarities with other proteins from the HEAT family without known connection to SMC-kleisins. This raises an important issue, and motivated the development of a more rigorous method for assessing relationships.

By definition, repeat proteins contain multiple copies of a homologous sequence motif. Assuming for the sake of argument that each repeat has a roughly similar probability of producing a significant alignment, then the more repeats a protein has, the greater the probability of it appearing as a significant hit in searches with homologous repeat proteins. Thus, when searching for relatives of the hawks, the presence of many proteins in the list of hits may not be indicative of close relationships, but rather a reflection of the fact that large HEAT proteins have more opportunities to produce significant alignments.

In order to deal with this issue, we reasoned that, if two repeat proteins are closely related via a single common ancestor, then they should each produce high-ranking alignments with the other, regardless of which was used as the query sequence (alignments are ranked in a fashion familiar to users of BLAST) In contrast, high-ranking alignments between highly diverged repeat proteins and generic proteins containing many canonical copies of the repeat motif should only appear when the diverged protein is used as a query. This is because the diverged protein will be attracted to large repeat proteins in a way that isn't true to the same extent in reverse.

Thus, for every 'seed' protein queried, I performed reciprocal searches using all proteins that produced significant alignments with that seed. Using the combined results from these searches, I generated a network, with each protein being represented by a node, and edges between them being weighted by the mean of the two alignment ranks. Thus, edges with high weights are broadly indicative of closer ancestry between two nodes than edges with lower weights. Finally, I clustered these networks using the MCL algorithm[148], which in our case revealed numerous well-defined sub-families of HEAT-related proteins. This method is summarised graphically in figure 5.2 below, and is also discussed in detail in methods section 7.2.4.

To initiate our network, I selected a semi-arbitrary set of candidate hawk proteins from budding yeast (*S. cerevisiae*), including known HEAT repeat proteins and those that we suspected might be related. After carrying out Amongst a large number of diverse clusters, all HEAT repeat proteins known to interact with the α-, β- and γ- kleisins[313,314] form a distinct, and highly robust cluster. Repeating this in humans and *S. pombe*, two clusters are formed, one containing SA/Scc3 proteins and a number of pseudogenes, and another containing all remaining condensin and cohesin hawks. The human network is shown below, with clusters represented by different colours (figure 5.3), and additional networks from yeast species are available in appendix figures A.14 and A.15.

It should be noted here that the three species networks are not independent from each other, but are complementary. This is a result of the fact that HHblits generates profile HMMs for each protein, which themselves are generated from multiple sequence alignments. Orthologous proteins from related species will therefore contain a considerable degree of overlap in the information content of their profile HMMs. However, replicating the networks in multiple species is useful as it allows us to look for hawks that may have been gained or lost in different species, such as Pds5A and Pds5B.

In order to validate the clustering of the hawks into a single group, I first carried out permutation tests by randomly shuffling the ranks of all alignments and regenerating the networks based on the newly assigned ranks. I repeated this process 10^6 times, and each time observed whether or not a cluster containing all the hawks, or all but the Scc3-related hawks was obtained. The results from

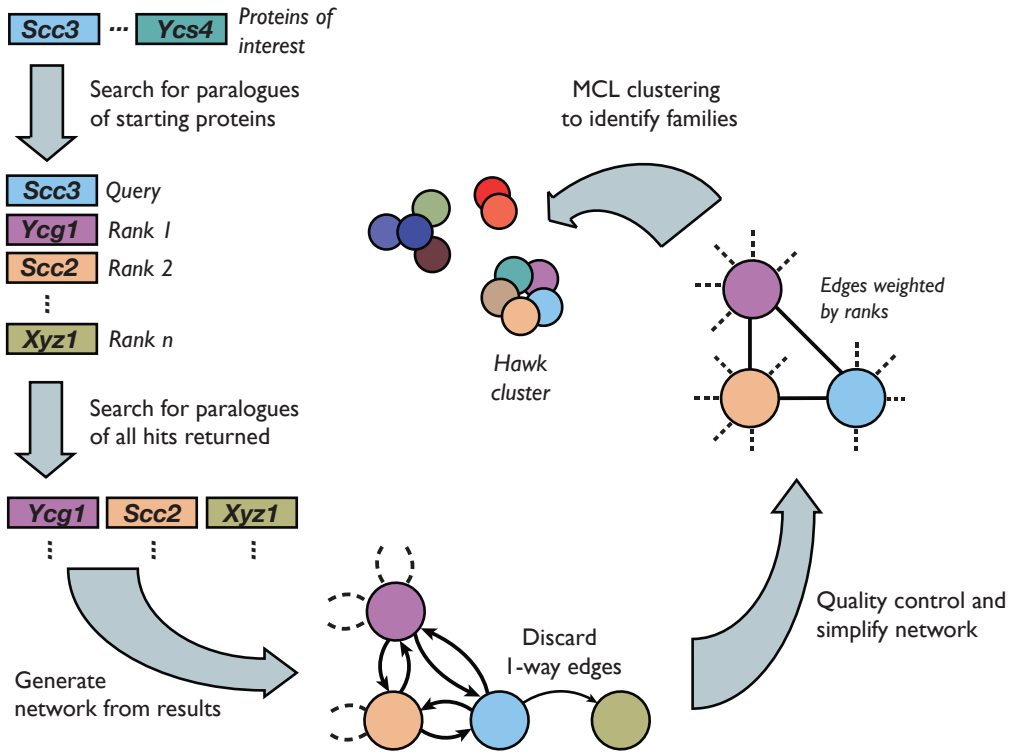


Figure 5.2.: Graphical summary of method
Computational pipeline for generation of homology networks in a single species. See methods 7.2.4 for further details. Adapted from figure S1, Wells et al.[3]

this showed the hawk clustering to be highly significant in all three networks (p-value $< 1 \times 10^{-6}$), with minor reductions in significance being achieved by allowing other proteins to cluster along with the hawks. As expected from the way in which alignment significance is calculated, I did not find a significant correlation between the length of alignment and its rank, indicating that clustering is not unduly affected by the size of the proteins involved.

I then sought to confirm that individual clusters in the networks contained useful biological information. Several known protein families were recapitulated in individual clusters, for example the Maestro family in the human network, whose members contain a shared HEAT-like repeat motif, and the clathrin adaptor family[327,334]. In addition, GO-term analysis demonstrated that many clusters are significantly enriched for similar biological processes (table 5.1). We are therefore confident that this method is robust, and that the clustering of hawks in a self-contained group is not an artefact.

5.2.2 Nse5 and Nse6 are erroneously annotated as containing HEAT repeats

Two proteins that are associated with the Smc5-6 complex (in *S. pombe*), Nse5 and Nse6, have been previously reported as containing HEAT repeats[321,337,338], analogous to the hawks. Our analysis fails to support these earlier findings. Both Nse5 and Nse6 are markedly shorter than typical HEAT proteins, at 388 and 522 residues respectively in *S. pombe*. In contrast the hawks are typically over

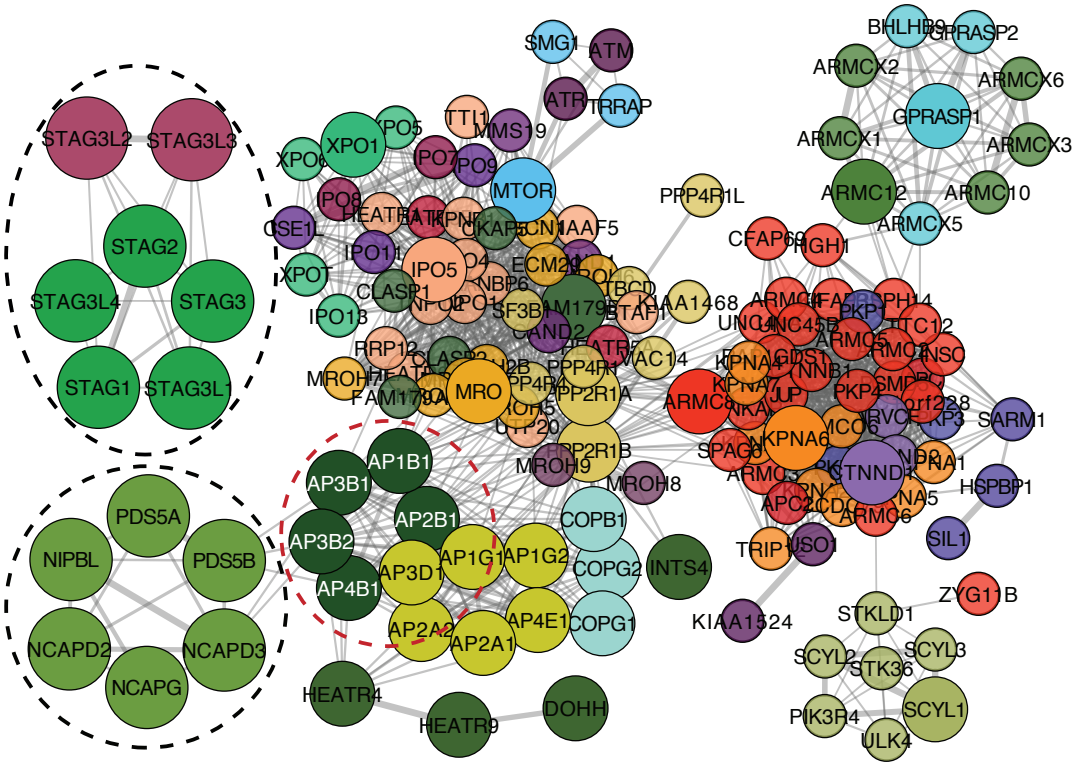


Figure 5.3.: Evolutionary relationships between HEAT repeat proteins captured by network clustering
Homo sapiens network. Different colours represent different clusters obtained by MCL and the hawk clusters are ringed in black. Thicker edges indicate higher mutual ranks and to aid visualisation only edges with a mean true positive probability $\geq$ 99.5% are shown. Subgraphs and individual nodes that become disconnected upon application of this threshold are not shown, with the exception of the STAG2 paralogues and associated pseudogenes; though closely related to the hawk cluster, they appear to have diverged significantly further than other members and cluster separately in the human and *S. pombe* networks. NCAPG2 resides in the PDS5 cluster but is not shown here due to edge weights not reaching the required threshold - all are above 97% but below 99.5%. Also note the clathrin adaptors - white labels, circled in red. Adapted from figure S2, Wells et al.[3]

1000 residues in length, and other HEAT proteins are often longer still. Though both proteins were included as seed sequences for the generation of our networks, neither clustered with the rest of the hawks, or indeed with anything else in the case of Nse5. On the basis of this initially unexpected finding, I manually inspected the results from the Nse5-6 HHblits searches, and found that neither had alignments meeting reasonable significance thresholds with either the hawks or indeed any other HEAT proteins.

We then searched the literature for the primary HEAT annotations of these two proteins. To our knowledge, the first paper to suggest that Nse6 contained HEAT repeats was Pebernard et al., 2006[339], which assigned the annotation on the basis of a result from a now obsolete structure prediction program, 3D-PSSM[340]. I therefore tried to replicate the finding with Phyre2[341], the successor program to 3D-PSSM, and found that the best-matched fold was a pair of alpha helices

Cluster	GO Description	GO-ID	P-val	Cluster %
8	sister chromatid segregation	819	1.32E-08	38.4
8	nuclear division	280	2.21E-08	53.8
7	protein amino acid phosphorylation	6468	8.36E-07	100
7	post-translational protein modification	43687	8.66E-06	100
3	NLS-bearing substrate import into nucleus	6607	4.75E-18	63.6
3	protein import into nucleus	6606	2.68E-15	72.7
2	protein transport	15031	1.07E-04	60
2	establishment of protein localization	45184	1.07E-04	60
1	response to indole-3-methanol	71680	2.25E-03	15.3
1	protein complex assembly	6461	4.83E-02	30.7

Table 5.1.: Sample GO-term enrichments
Exemplary GO-term enrichments from a selection of *H. sapiens network*. A complete list of cluster members can be found in appendix table A.1. P-values calculated with hypergeometric test and corrected for false discovery rate using the Benjamini-Hochberg method[335,336]

arising from an Integron cassette protein (PDB code: 3JRT). The Phyre2 score for this prediction was marked as being low confidence (56%) and the template protein has no similarity to typical HEAT repeat proteins.

Curiously, I did see one hit with HHblits between Nse6 and Cnd3 in *S. pombe* spanning a region approximately the length of 1 repeat (51 residues). However, this was ranked 8th out of a total of 10 alignments (typical proteins yield tens to hundreds), with a true positive probability of 9.8%, an expect value of 340 and p-value of 0.067. This alignment disappeared under different search parameters (both more and less sensitive), was not detected in humans, and budding yeast does not appear to have any orthologues of Nse6. This clearly does not meet reasonable significance thresholds, but it is hard not to be intrigued. Whilst it could be indicative of either convergent or highly divergent evolution, it seems most likely that it that it is a spurious result arising from multiple testing.

With respect to the Nse5 annotation, I was unable to find any published evidence for it containing repeats in either the Pebernard paper or others, and am unsure as to where the annotation originally came from. Using HHRepID[342], I then searched for any evidence of repeats, HEAT or otherwise, but found none under a range of different parameters and sensitivities. However, having ruled out the possibility of Nse5-6 being bona-fide hawks, an important result arises, namely that condensins and cohesins have hawks, but have lost the kites. In contrast, Smc5-6 is alone amongst eukaryotic Smc-kleisin complexes in retaining kites (the Nse1/3 subunits), but lacking hawks.

For the avoidance of further confusion, I would also note that *S. cerevisiae*'s Kre29, though likely performing a similar role to *S. pombe*'s Nse6[343], shows no indication of being evolutionarily related. In humans however Slf2, although approximately twice the length, does produce significant alignments through its C-terminal end with Nse6 (true positive probability: 97.8%, expect value: 2.81×10^{-6}). Finally, The cohesin regulator Scc4/MAU2 interacts with Scc2 and contains tetratricopeptide repeats, which are superficially similar in structure to HEATs. We initially considered

the possibility that it might be related to the hawks; ultimately however, Scc4 (PDB code: 4XDN) is structurally very different, with much tighter curvature and a different alpha-helix layout. In terms of sequence, there is no apparent homology between Scc4 and any of the hawks, and thus we feel confident in saying that Scc4 is unrelated to the hawks.

5.2.3 Evolutionary origin of the hawk family

Having demonstrated the close evolutionary relationship between hawks, we next looked to a possible origin for the family. Searches of sequence databases revealed orthologues in a set of species collectively accounting for all of the eukaryotic family tree (table 5.2), strongly suggesting that the last eukaryotic common ancestor (LECA) contained at least one member of the family. It is known that several archaea species contain proteins with PBS lyase HEAT-like repeat domains[344], and I therefore decided to look for repeats in the recently discovered lokiarchaeaota, which is a member of the Asgard archaeal superphylum, members of which are amongst the closest relatives to the eukaryotes[345,346]. Although several significant hits were found, based on sequence annotations and simple size comparisons it seems almost certain that these proteins are not directly related to the hawks. Interestingly however, when I integrated them into the existing networks, they predominantly clustered with the clathrin adaptor and coatomer families, and the former of these appears to be amongst the most closely related families to the hawks (figures 5.3, A.17, A.18). The close relationship between hawks and clathrin adaptors is also consistent with results from the paper originally identifying HEAT repeats in hawk proteins[329].

5.2.4 Structural support for a common ancestor of hawks

Finally, I used recently published Pds5B[350] (PDB ID: 4PJU) and SA2[351] (5HDT) structures to further validate our findings from sequence analysis. Both structures share significant similarity in terms of overall morphology and their known interaction partners, Rad21 and Wapl. Furthermore, as the abridged version of this work was undergoing peer review[3] (see appendix A.2), the structures of Scc2 from *Chaetonium thermophilum*[315] and *Eremothecium gossypii*[352] (5T8V and 5ME3) were released, and they too shares the characteristic S-shape of Pds5B and SA2, which differs noticeably from most other HEAT proteins.

Pds5B and its orthologues in yeast contain a large, centrally located insertion or deletion that manifests as a large protrusion from the side of the structure (figure A.16). When I split the structure on either side of this alignment, I found that the two parts on either side aligned very well to the SA2 structure, with TM-scores significantly higher that expected for unrelated sequences (figure 5.4A), an observation that was also noted independently by Lee et al.[353]. Although a degree of skepticism is warranted due to the fact that tandem structural repeats are probably more likely to produce significant structural alignments than more traditional structures, the similarities here are convincing. Further supporting this, I generated multiple sequence alignments of metazoan orthologues of Pds5B and SA2 and used these to map sequence conservation onto the surfaces of the two structures, revealing broadly similar patterns of conservation that correspond to the binding regions of Rad21 and Wapl (figure 5.4B)

Supergroup	Group	Species	Hawks
Archaeplastida	Viridiplantae	*Arabidopsis thaliana*	yes
Archaeplastida	Rhodophyta	*Galderia sulphuraria*	yes
Archaeplastida	Glaucophyta	*Cynanophora paradoxa*	**no**
Excavata	Diplomonadida	*Giardia lamblia*	**no**
Excavata	Parabasalia	*Trichomonas vaginalis*	yes
Excavata	Kinetoplastida	*Trypanosoma cruzi*	yes
Opisthokonta	Fungi	*Saccharomyces cerevisae*	yes
Opisthokonta	Metazoa	*Homo sapiens*	yes
Amoebozoa	Mycetozoa	*Dictyolstelium discoideum*	yes
Amoebozoa	Lobosa	*Naegleria gruberi*	yes
SAR	Rhizharia	*Plasmodiophora brassicae*	yes
SAR	Stramenopiles	*Aphanomyces invadans*	yes
SAR	Alveolata	*Oxytricha trifallax*	yes
Unknown	Haptophyta	*Emiliana huxleyi*	yes
Unknown	Cryptophyta	*Guillardia theta*	yes

Table 5.2.: Sample of eukaryotic supergroups with hawk orthologues
Sequences were searched for manually using PSI-BLAST and HMMER[333,347]. *G. lamblia* (4th from top) is an intriguing case since it possesses Smc orthologues and carries successfully carries out chromosome condensation and segregation[348], but apparently lacks an obvious Scc1/Rad21 kleisin orthologue[348,349]. In light of this, it is less surprising that it does not possess hawks either, since all of those associated with cohesin (and probably of condensin) are known to bind to the kleisin subunit. Of course, the obvious question arising from this is how condensin and cohesin function in this organism?

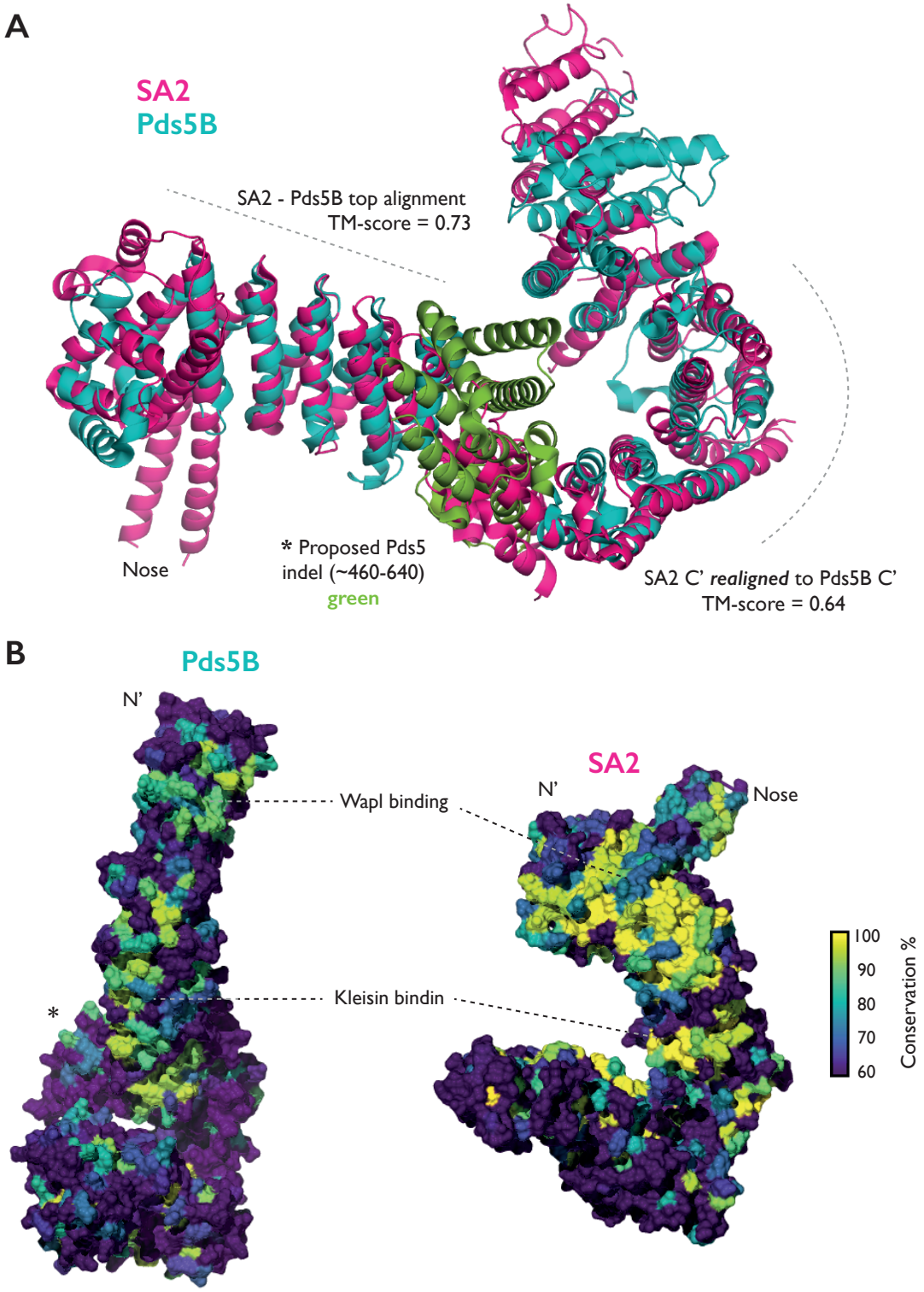


Figure 5.4.: Structural similarities between human hawks Pds5B and SA2
(A) Structural alignment of Pds5B and SA2. Overall, the two structures align moderately well, but are disrupted by a large indel in the middle of Pds5B, which appears to be conserved across all Pds5 orthologues. When splitting the Pds5B structure into three pieces at the start and end of this indel, the two terminal pieces align much better to SA2. (B) Both structures bind the alpha-kleisin Rad21 and Wapl. They have similar patterns of conservation along the convex face of their structures, corresponding to the binding location of Rad21. Similarly, towards their N-termini they both have Wapl binding sites. Adapted from figures 1 and S2, Wells et al.[3]

5.3 DISCUSSION

The results of this work lead to a number of conclusions, giving rise to the tentative model shown in figure 5.5. Firstly, we are confident that the hawk family is descended from a common ancestor. Although they have diverged too far from each other to build an accurate phylogenetic tree, their robust clustering into a single group, along with their structural and functional similarity strongly suggests that they are monophyletic. This explanation is more parsimonious than the alternative, in which convergent evolution lead to multiple HEAT repeat proteins being recruited independently to the Smc-kleisins for similar functions. Nonetheless, it will be of great interest to see the structures of the hawks associated with the condensins, as these will provide a more complete picture than we currently have with the cohesin hawks alone.

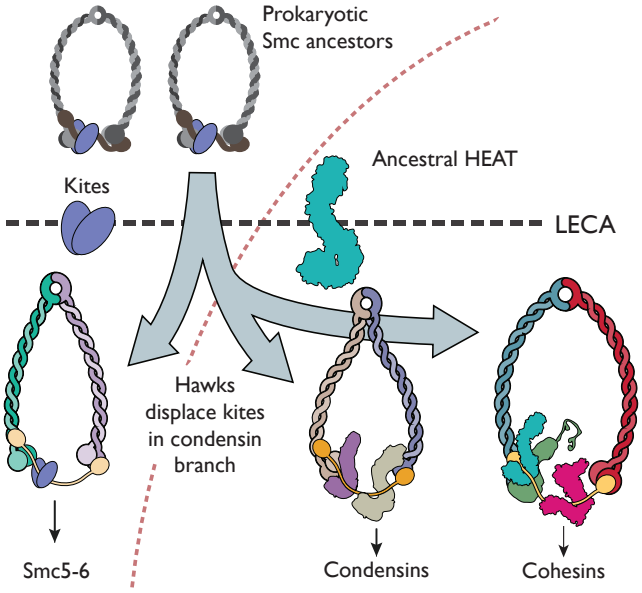


Figure 5.5. Proposed origin of eukaryotic SMC-kleisins
Kite proteins are found in many bacterial and archaeal SMC-kleisins, and also the eukaryotic Smc5-6. We suggest that, very early in eukaryotic history, an ancestral HEAT repeat protein related to the modern-day clathrin adaptors became associated with the ancestral eukaryotic SMC complex. Over time, subsequent duplications of this protein displaced the kites and lead to the condensin/cohesin split. Adapted from figure 1, Wells et al.[3]

It also seems likely that previous papers that had annotated Nse5 and Nse6 as containing HEAT repeats are incorrect on this point. I was unable to replicate the findings first reported by Pebernard et al.[339], using a number of more powerful approaches. Since Nse5 and Nse6 are not hawks, this implies that Smc5-6 is unique in having retained the kite proteins, and has either lost or never had hawks. Thus it seems that the gain of ancestral hawk proteins via successive gene duplications may have been the decisive event that initiated the evolution of present day condensins and cohesins.

The fact that various hawk homologues are found across all extant eukaryotic branches suggests that they arose around the time of LECA, if not earlier. It does not seem likely that archaeal species possessed them, though the similarity between HEAT proteins in lokiarchaeota, the clathrin adaptors and the hawks is also indicative of their early evolution. It is interesting to speculate about how these results fit into other studies on the origin of the nucleus[354,355], and the degree to which linear chromosome condensation does or does not overlap with the evolution of nuclear structure.

Finally, there has been considerable debate about whether or not bacterial homologues of the SMC-kleisin complexes are 'bacterial condensins', as they are commonly referred to as[314,356,357]. Our analysis demonstrates that there are definite compositional differences between those SMC-

kleisins that contain kite proteins, including prokaryotic SMC-kleisins and Smc5-6, and those with hawks - condensin and cohesin - on the other. Nonetheless, there are consistent reports of bacterial SMC-kleisins being involved in behaviour highly similar to that of eukaryotic condensin[320,356]. This suggests that there may at least be functional overlap between prokaryotic and eukaryotic condensins, and work is currently underway to reveal the structural and functional details of the condensin hawks.

6 CONCLUSION

6.1 INSIGHTS INTO THE NATURE OF PROTEIN COMPLEXES

In this work I have attempted to develop our understanding of the mechanisms by which the cell regulates protein complex assembly, and of the implications this process has for the proteome as a whole. After a literature review covering the currently available methods that can be used to study protein complexes, I began by describing a study of gene order in prokaryotic operons, demonstrating that bacterial gene order matches the assembly order of protein complexes. This was followed by work on protein degradation kinetics in mammalian cells, which revealed novel insights into the nature of eukaryotic protein complex assembly. Using a simple model developed in that chapter, I then showed how rapid degradation of excess heteromeric subunits could explain the phenomenon of protein attenuation in aneuploid cells. Finally, I described the Hawk proteins - an important family of condensin and cohesin regulators descended from an ancestral HEAT repeat protein.

Aside from this last excursion into evolution, the primary focus of this thesis has been on regulatory mechanisms governing the assembly of protein complexes, both in prokaryotes and eukaryotes. Chapter 2 focused on prokaryotes - bacteria in particular - demonstrating that gene order in operons is under evolutionary selection pressure to match the assembly order of heteromeric protein complexes. Earlier work had shown that gene fusions were more likely to be fixed if they preserved the order of assembly[104], but since these are relatively rare events, the finding that operon gene order is also constrained provides stronger evidence for the biological importance of ordered assembly. It is now clear that significant fitness benefits can be achieved via mechanisms that guide assembly down thermodynamically favourable pathways.

A second important implication from this study relates to the location of assembly in bacteria. In order for gene order to have a beneficial effect on assembly efficiency, assembly must be taking place very close to the site of translation. This implication has been directly supported by an experimental paper in which showed that operon-encoded luciferase subunits associated co-translationally in *E. coli*[44]. Furthermore, numerous empirical reports and theoretical arguments in the literature indicate that this is not a phenomena unique to prokaryotic operons[264,265].

Eukaryotic protein complexes do not benefit from being encoded in operons, and therefore other mechanisms must have evolved that allow robust assembly of complexes. An important distinction between prokaryotes and eukaryotes is that the latter typically have smaller effective population sizes - this places constraints on the ability of selection to leverage small fitness benefits[358,359].

However, protein complexes are such a fundamental part of the cell that the process of assembly is certainly not left entirely to chance. One such mechanism with the power to facilitate assembly of eukaryotic complexes is pointed to in chapter 3 - a study of protein degradation kinetics. This study revealed that many proteins are degraded non-exponentially, and this appears to be a result of rapid degradation of excess protein complex subunits.

The key finding from this work is that eukaryotic protein complexes - in contrast to prokaryotic ones[227] - are not expressed at levels in accordance with their stoichiometries. Instead, eukaryotic protein complex assembly appear to assemble around core subunits that are overexpressed relative to their stoichiometric requirements, with excess protein being rapidly degraded (figure 6.1). These are fundamentally different mechanisms, and reflect the differences in genome architecture between these domains of life.

A consequence of the behaviour of eukaryotic heteromers during assembly can be seen in aneuploid cells. Specifically, in the attenuation of proteins that occurs in response to increasing chromosomal copy number. From collaborative work carried out with others[2] and my own later investigations on the topic (chapter 4), it now seems clear that this phenomenon is a direct result of degradation of excess protein subunits, as has been suggested previously[305,306,360]. Having said that, there are still questions surrounding the degree to which protein aggregation occurs in aneuploid cells, and how much this confounds observations of attenuation.

Finally, whilst the characterisation of the Hawk family of condensin and cohesin regulators does not directly relate to assembly, it does shed light on the evolution of a fascinating group of protein complexes, namely the SMC-kleisins. The fact that the hawks appear to originate very early in eukaryotic history suggests that they played an important role in the evolution of condensin and cohesin's present-day functions, and highlights more broadly the importance of subunit gain and loss in the evolution of protein complexes[105,361]. Enticingly, the relationship between the hawks and the clathrin adaptor proteins hints at a link with the proto-coatomer hypothesis[362,363], and raises questions about how closely intertwined the emergence of linear chromosome condensation was with the evolution of the nucleus.

6.2 QUESTIONS ARISING FROM THIS WORK

Many new - and some old - questions are suggested by this work. As an obvious starting point, whilst coexpression of protein complex subunits is ubiquitous, it is still not at all clear how this is achieved in eukaryotes. Although eukaryotic gene order is non-random (see review by Hurst et al.[364]), there is very little clustering of genes in the manner of operons, which might lead to co-regulation due to spatial proximity of active transcription factors. However, even in the absence of spatial clustering, it seems to be rare for protein complexes to be co-regulated by single transcription factors[365].

This is a question that, whilst especially pertinent to the assembly of protein complexes, is relevant to any biological processes that require coordinated gene expression. The failure of simple mechanisms to explain coexpression suggests that it is an emergent property of the transcription factor network, involving multiple transcription factors operating at different regulatory levels[366,367],

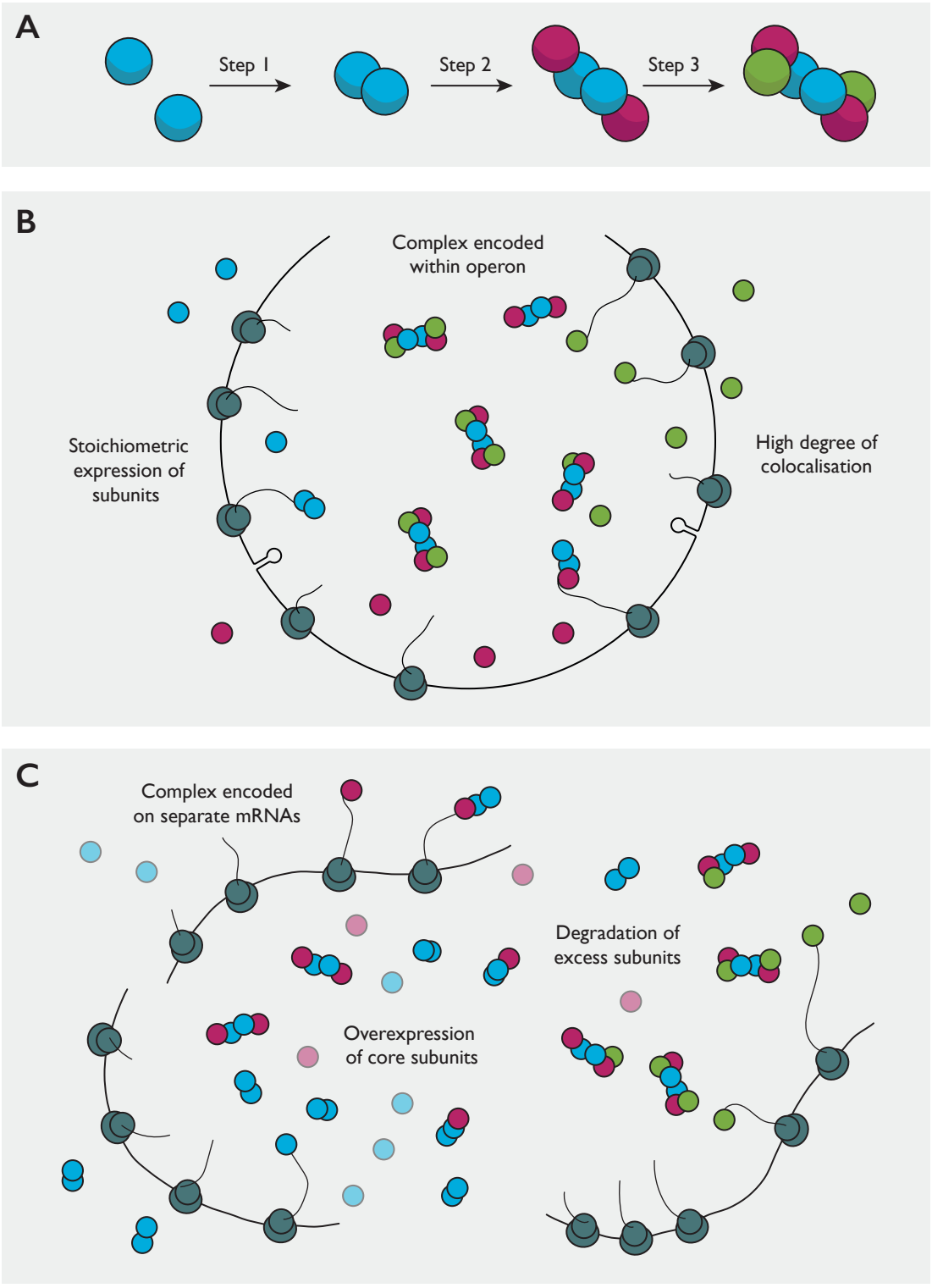


Figure 6.1.: Models of prokaryotic and eukaryotic heteromer assembly
(A) Thermodynamically favourable assembly pathway for a hypothetical protein complex. (B) Prokaryotes frequently encode heteromeric protein complexes within operons. Subunits from these are encoded in the same order as they assemble, and are expressed in proportions closely matching the stoichiometry of the complex. Furthermore, they benefit from a high degree of colocalisation afforded by operons, and cotranslational assembly is almost certainly commonplace. (C) Eukaryotes do not encode protein complexes within operons, but instead appear to compensate for the lack of co-regulation incurred by this by overexpressing core subunits and rapidly degrading the excess.

but explaining this will therefore require a deeper understanding of cellular transcription factor networks. Projects such as ENCODE have done much to further progress in this domain, but there are still fiery debates in the field, even about seemingly basic issues such as the size of the functional genome[368–370].

One point of interest from the work on gene order in operons is the implication that cotranslational assembly is likely the default for operon-encoded protein complexes. In the case of prokaryotic operons, it is easy to imagine this being true, but less so for eukaryotes, in which each protein is encoded on a separate mRNA. Nonetheless, there is considerable evidence for cotranslational assembly taking place in eukaryotes too, not only for homomeric complexes but also for heteromers - two reviews on this topic are included in appendix A.2. Still, there are plenty of unanswered questions on the topic of where in the cell assembly takes place. Without doubt, extensive subcellular localisation of proteins does occur, but whether or not this is sufficiently specific as to make an appreciable difference on assembly efficiency is still unknown.

With respect to the work on protein degradation kinetics, there are several leads that would be interesting to follow up on. The fact that both NED character and attenuation in aneuploid cells are reduced by repressing the proteasome motivates one to ask what role ubiquitination plays in the regulation of protein complex assembly and degradation? There is evidence that both ubiquitination and acetylation sites are enriched in heteromeric protein complexes[298,371], which hints at considerable complexity in the regulation of protein complex degradation[372]. An interesting project would be to ask whether or not such modifications affect the ability of protein complexes to assemble - one way of investigating this might be to look at whether there is any enrichment or depletion of modified lysine residues in the interface regions of protein subunits.

Also, whilst it seems likely that rapid degradation of excess subunits is of benefit partly because it reduces the likelihood of off-target interactions - the need for which is an important constraint on protein evolution[220] - there are more subtle mechanisms for controlling expression that could also be facilitated by this system. Under the model described in chapter 3, the simplest interpretation of a protein complex containing both ED and NED proteins is that the NED proteins appear as such because they are present in excess, and thus have both bound and unbound subunits, whereas the ED proteins are fully bound, hence the single degradation rate.

This implies that the ED subunits are limiting for the overall abundance of the complex, and therefore by controlling the expression of a small number of ED subunits, overall levels of the protein complex could be modulated relatively simply. Testing this hypothesis would be challenging, though not impossible, and would require changes in subunit expression to be monitored over extended time periods or in response to cell perturbations. Interestingly, some circumstantial evidence for this idea exists in a nice paper showing that many protein complexes involved in the yeast cell cycle consist of a mixture of constitutively and periodically expressed subunits[373].

There is still scope for further investigation into the attenuation of proteins in aneuploid cells. For example, although there seems to be a strong connection between NED and attenuation, the latter is a phenomena caused by fairly drastic changes to the karyotype of a cell. Duplicating a chromosome could reasonably be expected to increase protein aggregation in the cell, and from preliminary (unpublished) work this certainly seems to be the case, with members of protein complexes being

particularly affected - the question now is to try and understand what dictates why some proteins aggregate whilst others are degraded.

6.3 CLOSING REMARKS

Over the preceding pages I have presented work relating to numerous aspects of protein complexes and the mechanisms by which they assemble. Beyond these contributions, there is still an enormous amount to be discovered, and it is my hope that some of the value of this work will be in the new research it motivates. With the wealth of additional biological data that becomes available each year, there is plenty of opportunity to learn more, and if the many excellent papers I have read over the last three years are anything to go by, then protein complexes will continue to be a fascinating area of study for many years to come.

7 METHODS

7.1 INTRODUCTION

The methods provided in this chapter relate to work performed either exclusively by myself or in direct collaboration with others. Except where strictly necessary, any details relating to work that I was not closely involved with have not been included - these can instead be found in the online versions of the published papers included in appendix A.2. For the sake of accuracy I have only made minimal changes to descriptions as they appear in the published documents. Unless indicated otherwise, most of the sections that follow should therefore be considered as quotes from the published versions, edited for clarity. In these cases, I have obtained permission for reuse where required. Additional code relating to this work can be found at: https://github.com/jonwells90.

7.2 METHODS

7.2.1 Chapter 2: Operon gene order is optimised for ordered assembly of protein complexes

Protein structural datasets

We started with the full set of prokaryotic X-ray and electron microscopy structures in the PDB on June 12, 2014. We considered all heteromeric pairs of subunits from the same complex, defined as having at least two different protein chains of at least 30 residues each and mapping to different UniProt sequences from a single species. Complexes with known quaternary structure assignment errors[374] were excluded. Very large complexes with more than 24 subunits were excluded, because we have not shown that the assembly of these can be predicted accurately from their structures. Heteromeric subunit pairs were filtered for redundancy at the level of 50% sequence identity.

Mapping subunit pairs to operons

Operon datasets were downloaded from the DOOR2 database[375]. Relevant datasets were identified based on the species and strain of each gene pair. After converting GI numbers to UniProt accession identifiers in each dataset, the set of gene pairs was mapped to the operon data. Operons encoding both members of a pair were added to a reference dictionary, with the locus and directionality of each gene being used to arrange constituent genes in order of expression. In rare cases where the copy number of a gene within an operon was found to be greater than one, the position of the gene in the operon was taken to be that of the first copy to be encountered, reading in the 5' to 3' direction. The set was then filtered to remove redundant operons (i.e. identical operons from

similar strains or species). In total, 368 gene pairs (220 adjacent) were mapped to 192 unique operons, with the remaining 711 pairs being expressed in different transcriptional units. Similarly, we also mapped a set of 2,562 binary protein-protein interactions (IM-22059)[28] to the E. coli K-12 W3110 operons for the analysis displayed in figure 2.2C.

To assess whether the gene order of a pair was evolutionary conserved, we used the STRING v9.1 database[376]. For each pair, we manually assessed, using the STRING online interface, whether all occurrences of a given gene pair shared the same gene order within their local evolutionary group as defined in STRING. This is at the level of phylum (e.g. Firmicutes or Euryarchaeota) or class for proteobacteria. Gene pairs present across only a very limited evolutionary range (less than three genera) were not considered to be evolutionarily conserved. Gene pairs associated with evolutionary gene fusion events were identified as those sharing >40% sequence identity with a gene pair with evidence for fusion in STRING, similar to what has been done previously[104].

Abundance measurements

We mapped all protein complex subunits in our dataset against the sequences of prokaryotic proteins from PaxDB v4.0[249], selecting abundance measurements from proteins with >90% sequence identity to a subunit. The results in figures 2.1 and A.7 only use abundance measurements from *E. coli*, but the analyses in table 2.1 and figures A.1, A.5, and A.7 are repeated using combined measurements from all available prokaryotes and also using protein synthesis rates derived from ribosomal profiling[227].

Prediction of assembly pathways

Ordered protein complex assembly pathways were predicted in a manner very similar to what has been done previously[104]. First, the complex is considered in terms of its constituent subunits and the sizes of the interfaces that can be formed between any pair of subunits are calculated with AREAIMOL[60]. Our model assumes that assembly will proceed via formation of the largest possible interface. The process is then repeated by calculating all possible interfaces that could form between subunits and subcomplexes until the full complex is assembled. To define which of a pair of subunits assembles first and which assembles later, we consider the first step of assembly that brings the two subunits together within the same (sub)complex. Whichever subunit was part of a larger subcomplex prior to this step is defined as assembling first. For example, in the blue pathway in figure 2.4A, the blue subunit homodimerizes first and then interacts sequentially with the free red subunits, so the blue subunit is defined as assembling first. If, alternatively, the first step of assembly had been a heterodimerization between the blue and red subunits, then both subunits would be classified as assembling simultaneously. The source code for predicting assembly pathways from protein complex structures is available at https://github.com/marshlab/assembly-prediction.

7.2.2 Chapter 3: Degradation kinetics of proteins are explained by assembly of protein complexes

Protein structural dataset

Starting from the entire set of protein structures in the Protein Data Bank on 2016-02-24, we searched for all polypeptide chains with >70% sequence identity to a human or mouse gene. For genes that map to multiple chains, we selected a single chain sorting by sequence identity, then number of unique subunits in the complex, and then the number of atoms present in the chain. Pairwise interfaces were calculated between all pairs of subunits using AREAIMOL[60]. The normalised assembly order was calculated for all complexes, excluding those containing nucleic acid chains, by first predicting the (dis)assembly pathway as previously described using all the pairwise interfaces from each heteromeric complex[104] and implemented in the assembly-prediction package[1]. For subunits with multiple copies within a single complex, the average assembly order of each subunit type was considered. The normalised assembly order was defined so that the first subunit to assemble has a value of 0, the last has a value of 1, and the average value for all unique subunits in a complex is equal to 0.5.

Non-structural dataset

To complement the analysis of protein complexes of known structure, we also performed coexpression analyses on the non-redundant 'core' set of mammalian complexes from CORUM[215] (downloaded 2015-10-20). As CORUM preferentially uses human complexes in its non-redundant set, homologous mouse versions of each complex were generated by replacing each subunit/gene with its mouse counterpart, provided sequence identity was at least 70%. Sequence identities were calculated by collecting all mouse sequences for which NED/ED classifications were available and running BLAST on these against all genes in the CORUM core set. In cases where the identity of a subunit was ambiguous (as defined by CORUM), the first possible subunit for which homology data were available was selected.

For further validation of the tendency of NED proteins to be 'core' subunits, processed mass spectrometric data was acquired from the dataset published by Hein et al.[106]. Their definition of core stoichiometry signature was used, specifically those proteins matching the criteria of residing in the circle with radius: 1 (in $log10$ units), centred at: -0.5, 0 (abundance stoichiometry to interaction stoichiometry, see figure 3B[106]).

Coexpression analyses

Coexpression data were downloaded from COXPRESdb[377] (mouse dataset: Mmu.v13-01.G20959-S31479; human dataset: Hsa.v13-01.G20280-S73083). For each complex, the mean coexpression of each available subunit was calculated, using pairwise correlations with all other available subunits in the complex. Cases where fewer than three unique subunits were present in the complex were discarded, due to calculations of average coexpression for each protein being superficially identical.

eQTL analyses

Human protein complexes were downloaded from the PDB (2016-11-24) using a minimum sequence identity of 90% for all chains in order to exclude structures such as human-viral complexes. eQTL data for 53 tissues was acquired from GTEx (v6p release) and then mapped to the human complexes. In order to maximise the available data, tissue datasets were selected by using whichever tissue maximised the number of available eQTLs for each complex.

7.2.3 Chapter 4: Autosomal dosage compensation in aneuploid cells

Protein structural dataset

A large collection of *S. cerevisiae* complexes was compiled from the PDB by mapping any structures containing chains that mapped with at least 90% sequence identity to yeast proteins. Where overlapping subcomplexes existed, those with the greatest number of unique subunits were retained. Genes from these complexes were then mapped to the yeast aneuploidy dataset provided in Dephoure et al.[305]. Wherever genes mapped to multiple PDB structures, structures were selected on the basis of highest sequence identity. Assembly order predictions were calculated using the same assembly-prediction package as previously described[1].

Aggregation dataset

To briefly summarise, the experimental protocol for acquiring SILAC aggregation data is essentially the same as that described in Dephoure et al.[305], with the exception being that the whole cell lysate is centrifuged to extract the aggregated fraction separately. The data from the aggregate fraction in each experiment are initially normalised (by our collaborators) by subtracting the average of all $log2$ disomic ratios from single-copy genes (which should theoretically be zero) from each protein in the aggregate.

Further normalisation is carried out using mean centring. Specifically, each data point is rescaled such that the mean of all data points in each experiment (i.e. each disomic yeast strain) is equal to the overall mean disomic ratio of all experiments combined. This has the effect of reducing the variance attributable to batch effects - see figure A.13.

Normalised abundance calculations

Protein abundance data for yeast was acquired from PaxDB v4.0[249] and mapped to each gene in the structural dataset. These values were then normalised as follows:

$$f(x) = \log_2 x - \log_2 m$$

Where m is the median abundance of subunits within each complex. This normalisation procedure allows abundance differences to be measured in terms of log_2 fold-change.

Disorder predictions

Disorder predictions for the complete set of protein sequences from *S. cerevisiae* were generated using the command line version of IUPred Version 1.0[309,378], with the output being a disorder score for each residue in the sequences. An overall score for each protein was given my taking the mean disorder score across all residues.

7.2.4 Chapter 5: Hawk proteins: A paralogous family of eukaryotic SMC-kleisin regulators

Construction of homology networks

Proteome fasta files for *S. cerevisiae*, *S. pombe* and *H. sapiens* were downloaded from the UniProt reference proteomes databank[191] (2016-04) and HHSuite v.3.0.0 was compiled from source[332,379]. HHsuite databases were constructed as per the protocol described in the HH-suite manual (available at http://www.mpibpc.mpg.de/soeding or https://github.com/soedinglab/hh-suite), using the clustered uniprot20_2016_02 database. It should be noted that due to the fact that HHsuite databases are generated from large multiple sequence alignments for each protein, the resulting species databases are not independent. Orthologous proteins in each species will, by virtue of that fact, produce profile HMMs with significant overlap.

Seed sequences for putative members of the Hawk family were selected semi-arbitrarily for each species. Each seed was searched against the uniprot20 database using HHblits[332] (local alignment, two iterations). Predicted secondary structure was added to each MSA/profile HMM using PSIPRED[380]. The resulting profile HMMs were then searched against the relevant species-specific database using HHsearch (local alignment, single iteration, no pre-filter) to generate a list of at most 500 putative paralogues from each seed. In turn, each one of these sequences was subjected to the same procedure, producing a large set of nodes and edges, with nodes representing proteins and edges representing alignments between them, weighted by the rank of the alignment.

The resulting graph was filtered by removing edges arising from alignments with a length of less than 100 columns (accounting for the length of 2 HEAT repeats), an expect-value of greater than 0.01 (thus controlling the false-discovery rate) or a true positive probability of less than 15%. Edge weights were then normalised according to the following formula, such that the normalised rank $f(r)$ lies between 0.01 and 1.0, with 1.0 being the best possible mean rank and 0.01 the worst.

$$f(r) = \frac{1}{1 + \frac{99(r - r_{min})}{r_{max} - r_{min}}}, 1 \leq r \leq 500$$

At this stage, each edge has a direction, pointing from the protein used as a query sequence to the returned paralogous protein. As such, a given pair of nodes can be connected by either one edge or two - the former only being possible if a protein appeared exclusively in the second round of searches and was therefore not queried itself. In order to make the graph undirected, all nodes with a degree of less than 2 were discarded and the remaining edges between each pair of nodes combined and weighted by the geometric mean of normalised alignment ranks. Since the geometric mean is always lower than the arithmetic mean, this avoids giving too much weight to results from proteins with very few significant alignments.

Finally, clustering was carried out using the mcl algorithm with an inflation parameter I = 2.5 for all networks[148]. Initial network construction and parameter setting was performed on a fully-labelled *S. cerevisiae* network, but *S. pombe* and *H. sapiens* replicates were performed on blinded graphs, with genes in each cluster only being revealed after all filtering and cluster parameters had been fixed. GO term enrichment analysis was carried out using the Cytoscape BiNGO app, with GO 'Biological Process' annotations[336]. P-values were generated using the hypergeometric test and corrected for false discovery rate using the Benjamini-Hochberg method[335,336].

Homology network permutation tests

Assuming a null hypothesis under which alignment ranks contain no information about the relative likelihood of two proteins being related, a single control network was constructed for each species. This was generated from the observed network by randomising the edge weights between each pair of nodes. This was achieved by pre-filtering alignments as usual, but randomly assigning ranks. These were then normalised and averaged as for the observed network. Each random network was then clustered and each cluster tested for membership of Hawk proteins; specifically we ask: does there exist a cluster in the random graph containing exclusively those proteins from the largest Hawk cluster in the observed graph? This process was repeated 106 times for each species, and the resulting p-value calculated as the number of times the complete Hawk cluster was seen, divided by the number of trials.

Searching for lokiarchaeota HEAT repeat sequences

13 Lokiarchaeota proteins containing HEAT repeats were downloaded from the UniProt database; 9 on the basis of UniProt sequence annotations and an additional 4 proteins, including 2 fragments, on the basis of HHsuite searches and manual inspection. These sequences were searched against our human HHsuite database, and the resulting human sequences searched back against the lokiarchaeota database. A sub-graph was built using the same parameters as for the main eukaryote networks, leaving exactly 10 archaeal proteins remaining after quality control. The resulting set of edges was concatenated onto the human network and re-clustered.

Mapping of repeat domain boundaries

Sequences from *S. cerevisiae* hawks and clathrin adaptors were used to generate multiple sequence alignments with HHblits. Multiple sequence alignments were generated with the uniprot20_2016_02 database. These alignments were subsequently passed to the HHRepID web server (https://toolkit.tuebingen.mpg.de/hhrepid). The threshold p-value for assigning repeat domain families was kept at 0.01, and the threshold for suboptimal self-alignments was set to 0.1, also the default. The number of HHblits iterations was set to 0 since we had produced our own MSAs in the preceding step. Repeat predictions were collected from the HHRepID results with alignment stringencies between 0.0 and 0.3, depending on which value produced highest confidence predictions.

Structural alignments and conservation mapping

Structures for human Pds5B and SA-2 were downloaded from the PDB (5HDT[350] and 4PJU[351] respectively, 28.04.2016). Structures were aligned in PyMol using TM-align[381,382], both globally and locally by splitting SA-2 and Pds5B at residues L436 and Y462 respectively and realigning each half. Conservation mapping was performed using multiple sequence alignments generated as follows: For Pds5B and SA-2, 1000 metazoan sequences for each were retrieved from the NCBI non-redundant sequence database using blastp, then clustered to 90% sequence identity with usearch[383,384]. The remaining sequences were then aligned in forward and reverse directions with MAFFT, MUSCLE and GlProbs, with a final composite MSA being generated with MergeAlign[385–388]. Finally, these were mapped onto the PDB structures in Chimera[389].

Analysis of putative Nse5 and Nse6 HEATS

Specific searches for HEAT-containing Nse5 and Nse6 homologues were carried out with the same parameters as for the main network – HHblits with 2 iterations to generate profile HMMs, followed by HHsearch to find significant alignments in the three main species datasets. Kre29 was used in place of Nse6 for S. cerevisiae, and Slf2 for Human. Subsequent searches using HHblits/HHsearch were carried out with more iterations for the HHblits step – this increases sensitivity but at the cost of accuracy in determining relative rank of alignments. Additional searches were performed in a wider variety of species using the proteome datasets available on the HHSuite webserver. Next, HHRepID[342] was used to try and detect repeats within Nse5-6 themselves (as opposed to HEAT containing homologues). As before, human Slf2 was also checked, as was Kre29. Iterations ranging from 3-8 were used to generate the profile HMMs, thus spanning a wide range of sensitivities.

Finally, a literature search was performed to try and identify the published evidence for the Nse5-6 HEAT annotations. On the basis of evidence for HEATs in Nse6 presented by Pebernard et al.[339], we attempted to replicate their finding using the structural prediction server 3D-PSSM, which is now obsolete[340]. Following this, we used the Phyre2 web server[341] with the Nse6 sequence (UniProt id - O13688) using default settings.

A APPENDICES

A.1 SUPPLEMENTARY INFORMATION

A.1.1 Chapter 2: Operon gene order is optimised for ordered assembly of protein complexes

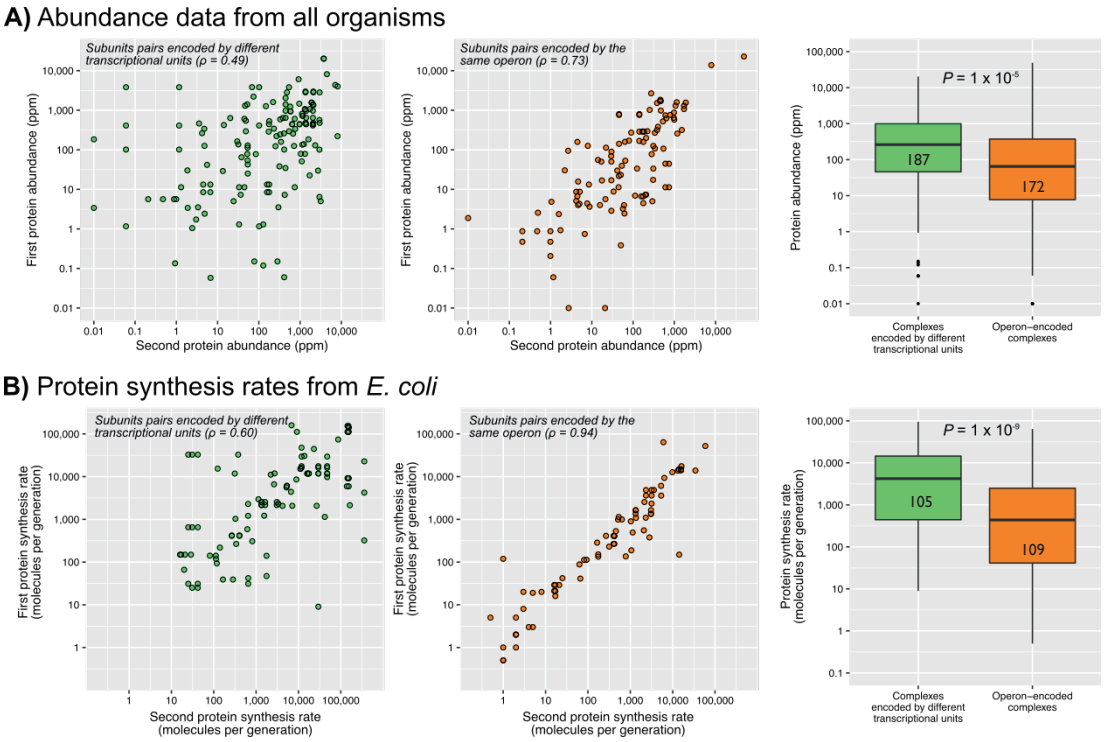


Figure A.1.: Additional comparisons of subunits pairs encoded in the same vs. different transcriptional units

(A) This figure shows the results from the same analysis as figure 2.1 B-C but using PaxDB abundance data for all organisms. The correlations between those pairs encoded in different transcriptional units is significant (p-value = 0.004), and calculated as figure 2.1 (B) Same as figure 2.1B-C but using protein synthesis rates from ribosomal profiling data[227]. P-value < 10^{-5}. Adapted from figure S1, Wells et al.[1]

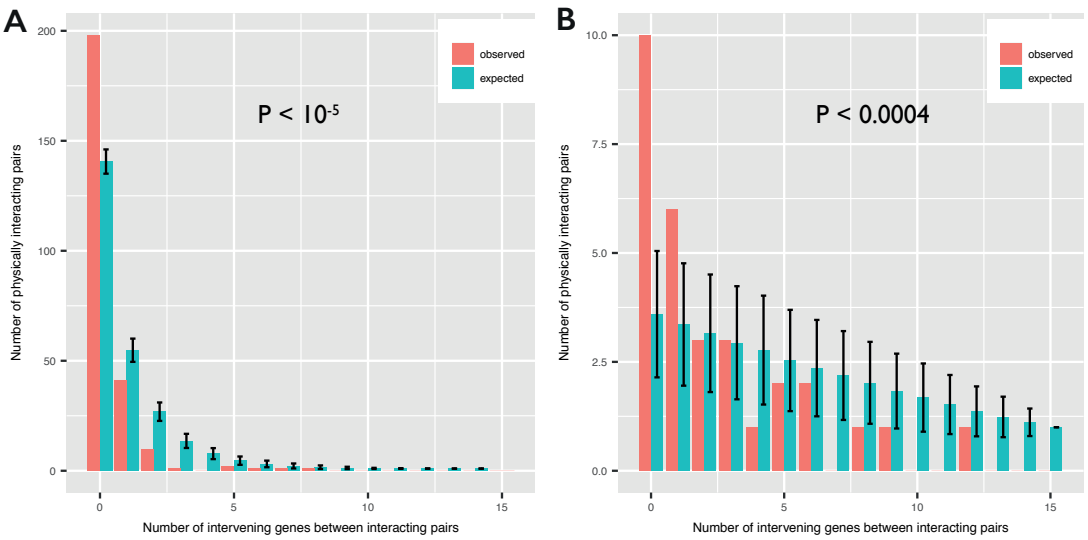


Figure A.2.: Relationship between gene pair proximity and likelihood of physical interaction, controlling for *nqo* operon

Panels (A) and (B) relate to the *nqo* operon from *Thermus thermophilus*, which encodes respiratory chain complex 1. Due to its size (17 genes), it accounts for more than half of non-adjacent gene pairs in our dataset (78/148). Within both the dataset when excluding it this operon and within the operon by itself (B), the observed number of interacting genes is higher than expected by chance. P-values calculated as for figure 2.3. Adapted from figure S2, Wells et al.[1]

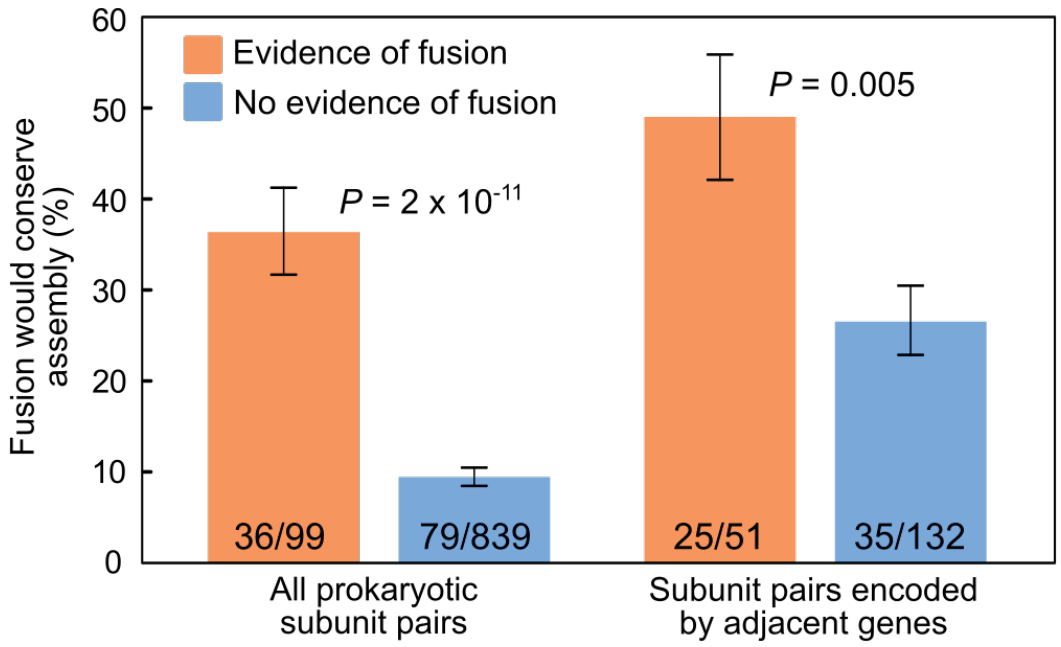


Figure A.3.: Gene fusion events conserve assembly order in adjacent gene pairs

Error bars represent Wilson 68% binomial confidence intervals and p-values were calculated with Fisher's exact test. Adapted from figure S3, Wells et al.[1]

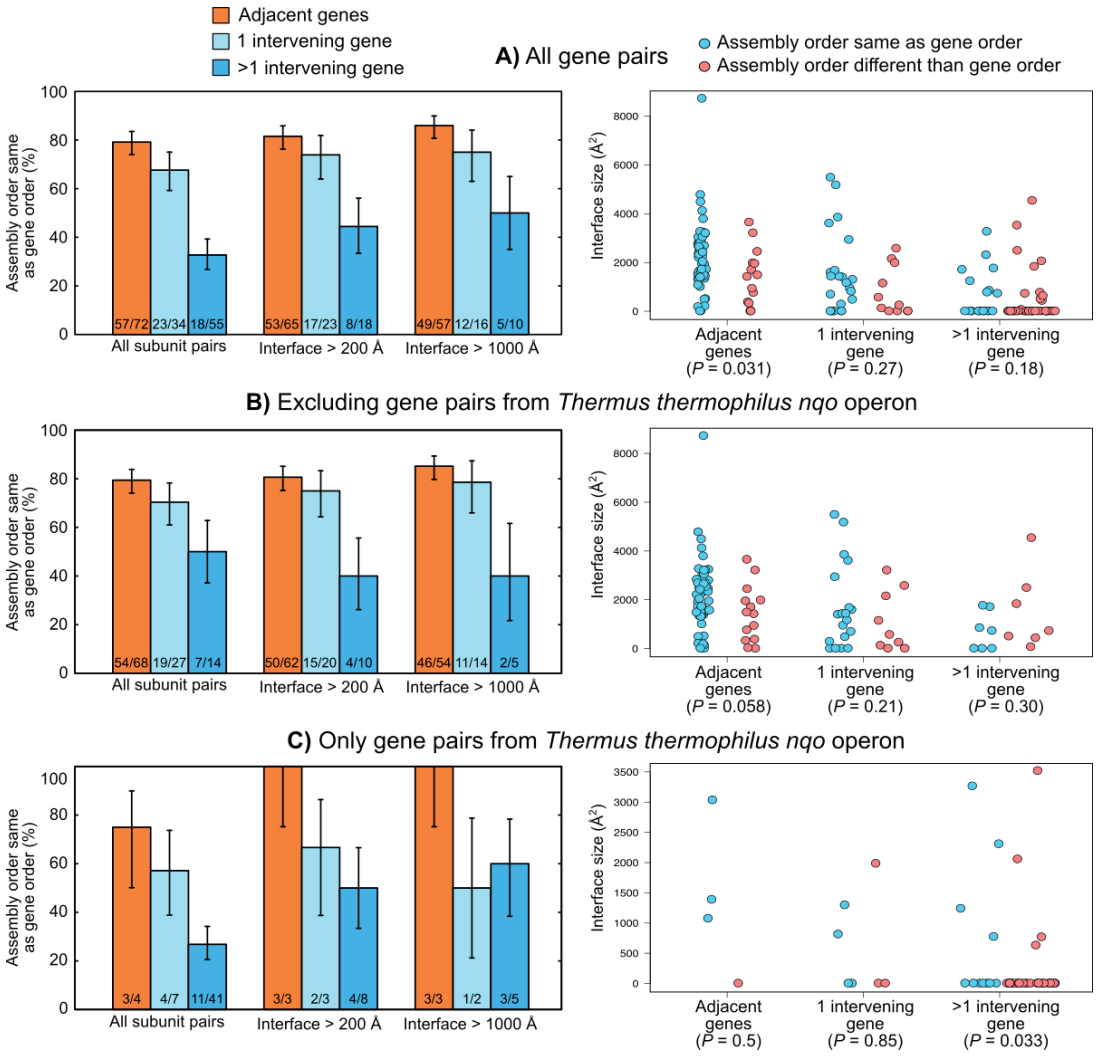


Figure A.4.: Comparison of gene order, assembly order and interface size for adjacent and non-adjacent gene pairs

Plots on the left describe the percentage of gene pairs for which assembly order matches gene order, split into adjacent pairs, those separated by a single intervening gene, and those separated by more than 1 gene (only genes with evolutionarily conserved order are shown.) Error bars are 68% Wilson binomial confidence intervals. On the right, plots show the distribution of interface sizes for interacting pairs where gene order matches or doesn't match assembly order. P-values calculated with Wilcoxon rank-sum tests. Adapted from figure S4, Wells et al.[1]

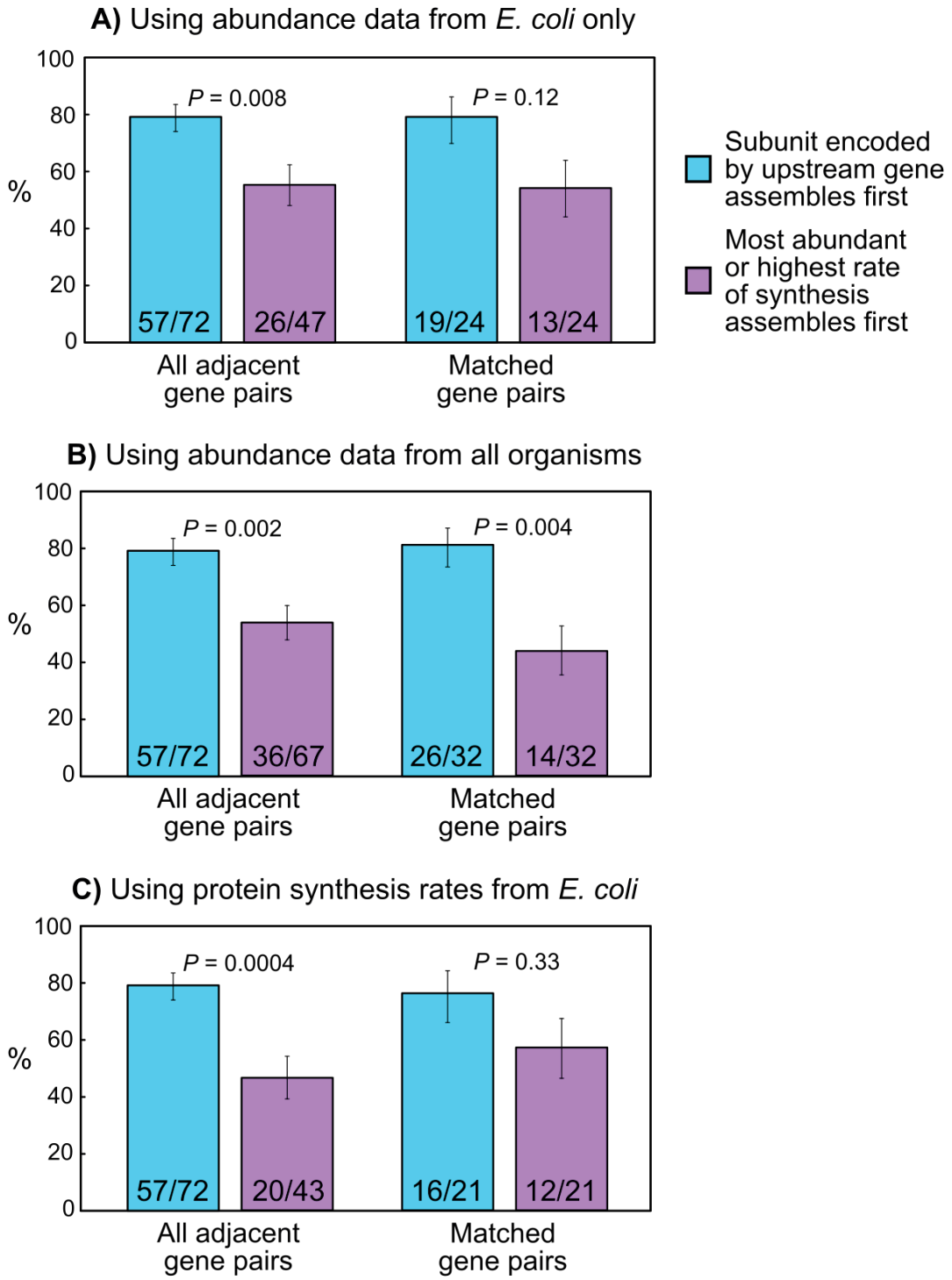


Figure A.5.: Gene order is a better predictor of assembly order than protein abundance
All gene pairs are those where gene order is conserved, error bars are 68% Wilson binomial confidence intervals and p-values are Fisher's exact test. Adapted from figure S5, Wells et al.[1]

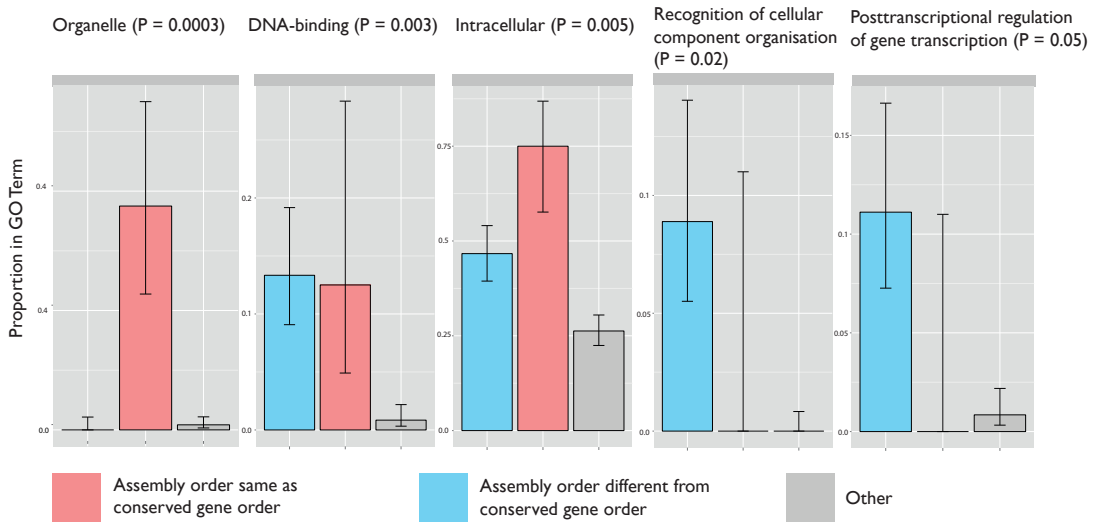


Figure A.6.: Enrichment analysis of gene ontology terms for gene pairs in which assembly order does not match gene order

Top five significant, non-redundant GO term enrichments for gene pairs in our dataset. In the above plots, 'Other' refers to cases where gene order is not conserved or where there is no well-defined assembly order.GO terms were filtered for redundancy, with terms appearing together in more than 50% of proteins in the GOA database, then only the most significant term was included in the non-redundant set. Error bars are 68% Wilson binomial confidence intervals and p-values were calculated using approximations of Fisher's exact test, based on 2×10^6 Monte Carlo iterations. The apparent enrichment for 'organelle' stems from just three complexes, and is thus probably not meaningful. Adapted from figure S6, Wells et al.[1]

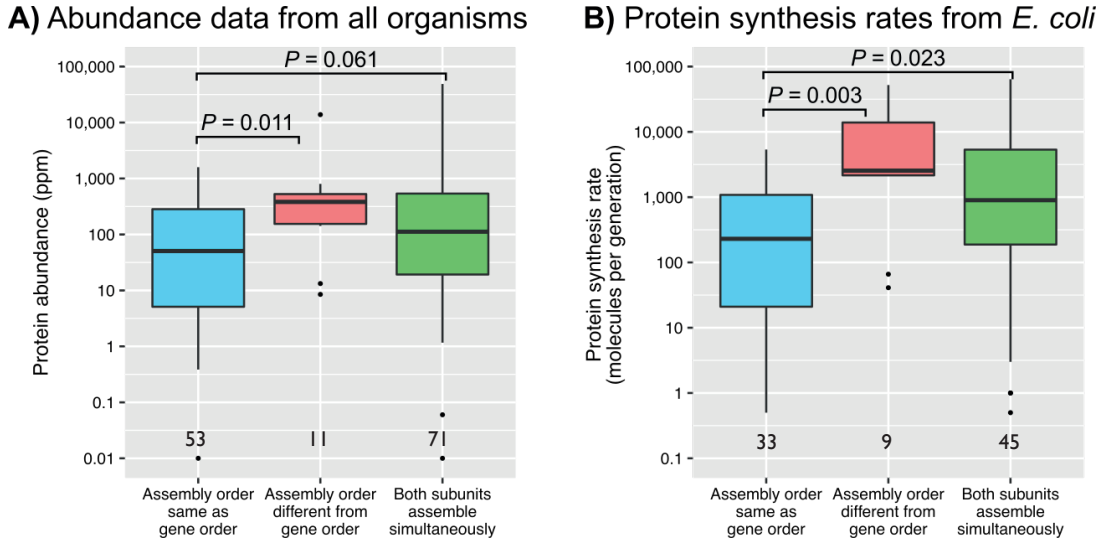


Figure A.7.: Additional comparisons of protein abundance for pairs where gene order matches assembly order and vice versa

These plots are the same as the analysis in figure 2.5, but using abundance data from all organisms or absolute protein synthesis rates. Adapted from figure S7, Wells et al.[1]

A.1.2 Chapter 3: Degradation kinetics of proteins are explained by assembly of protein complexes

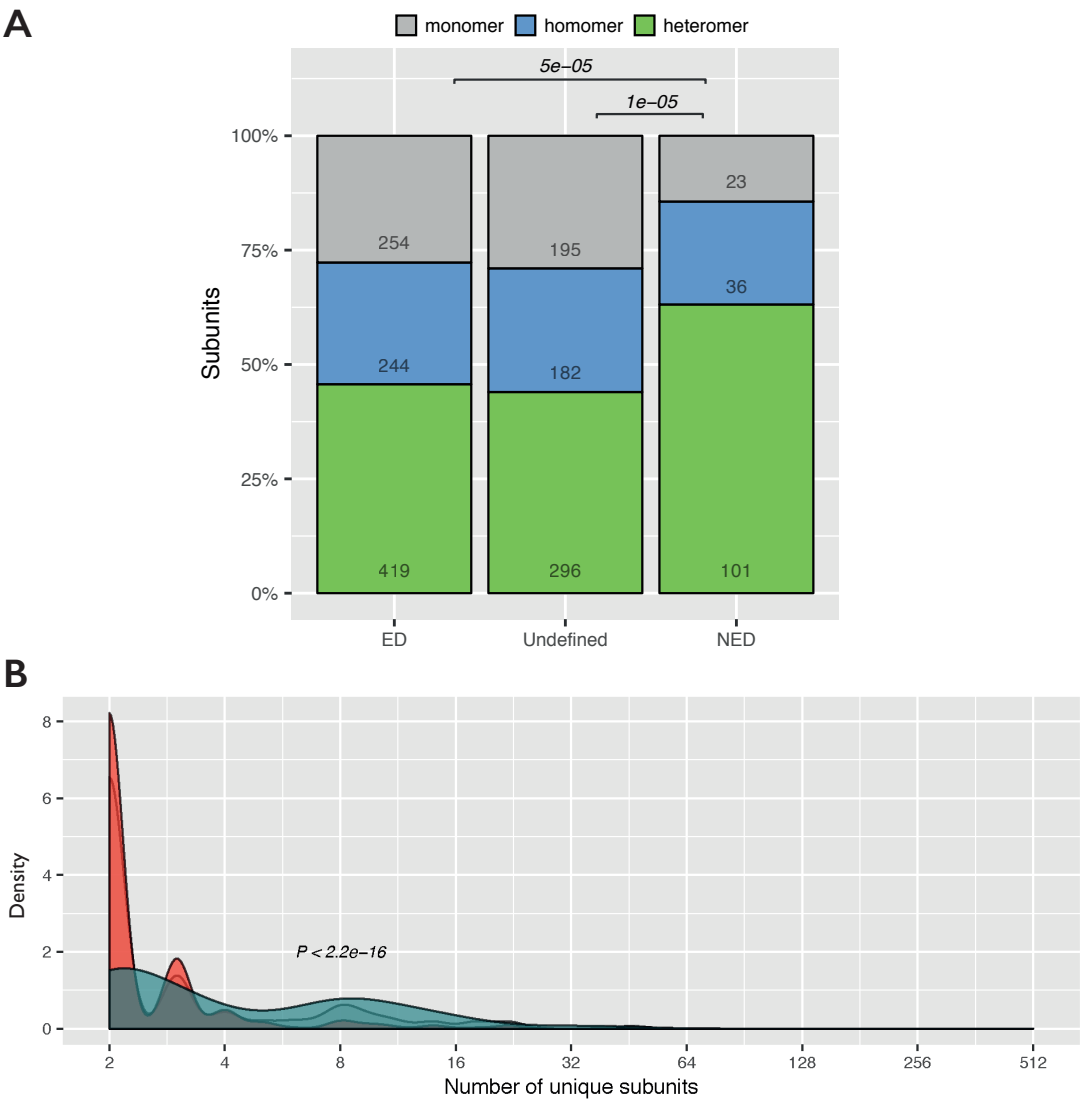


Figure A.8.: Enrichment of NED proteins in heteromers is independent of the presence of ribosomes
Since ribosomal subunits are prevalent in our dataset and are known to be degraded rapidly when in excess[219,268], I repeated the analyses in figure 3.4A-B, again finding highly significant differences.

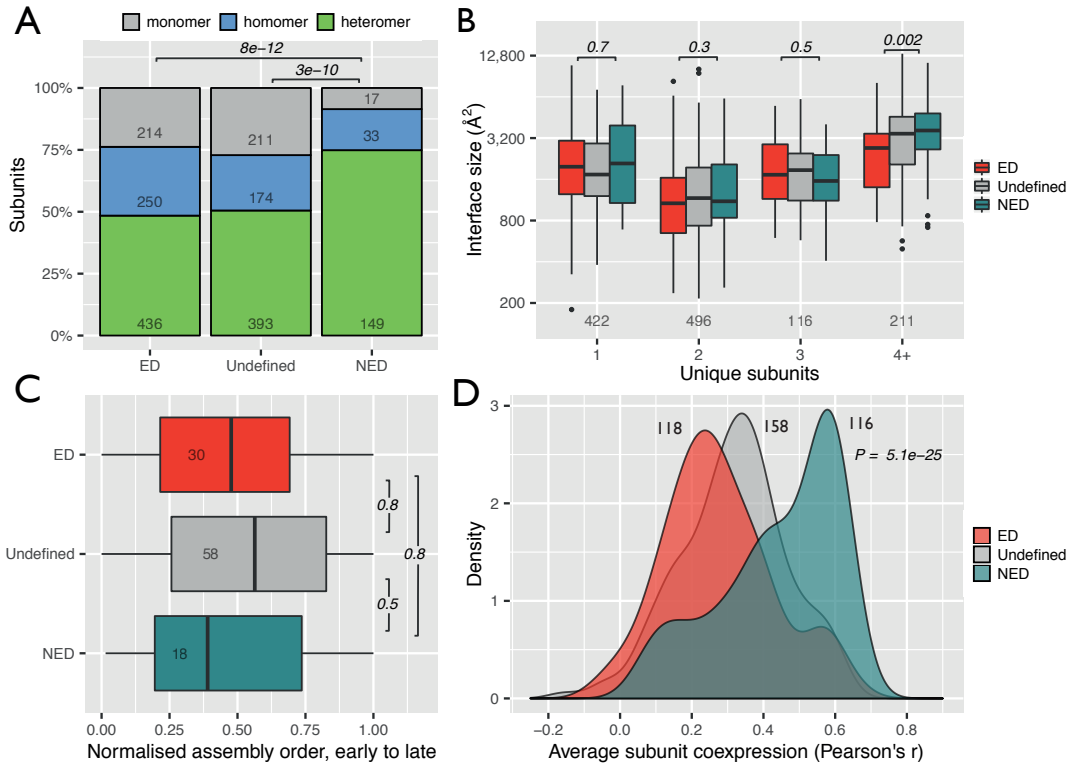


Figure A.9.: Non-exponentially degraded proteins are common - human
Replicated version of figure 3.4A-D using human data generated from RPE1 cells.

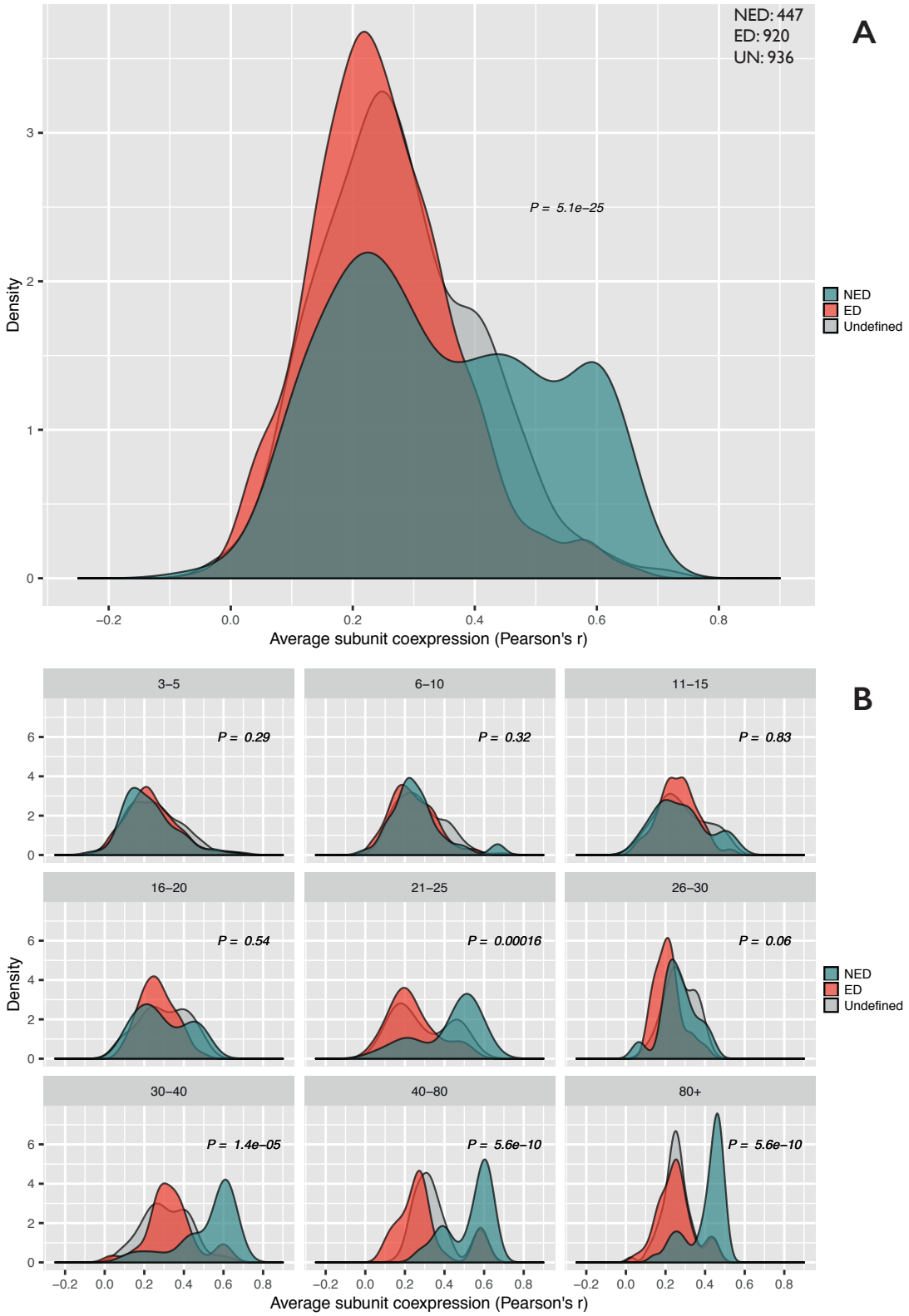


Figure A.10.: Increased NED protein coexpression is not unique to structural data - human
Replicated version of figure 3.5 using human data generated from RPE1 cells.

A.1.3 Chapter 4: Autosomal dosage compensation in aneuploid cells

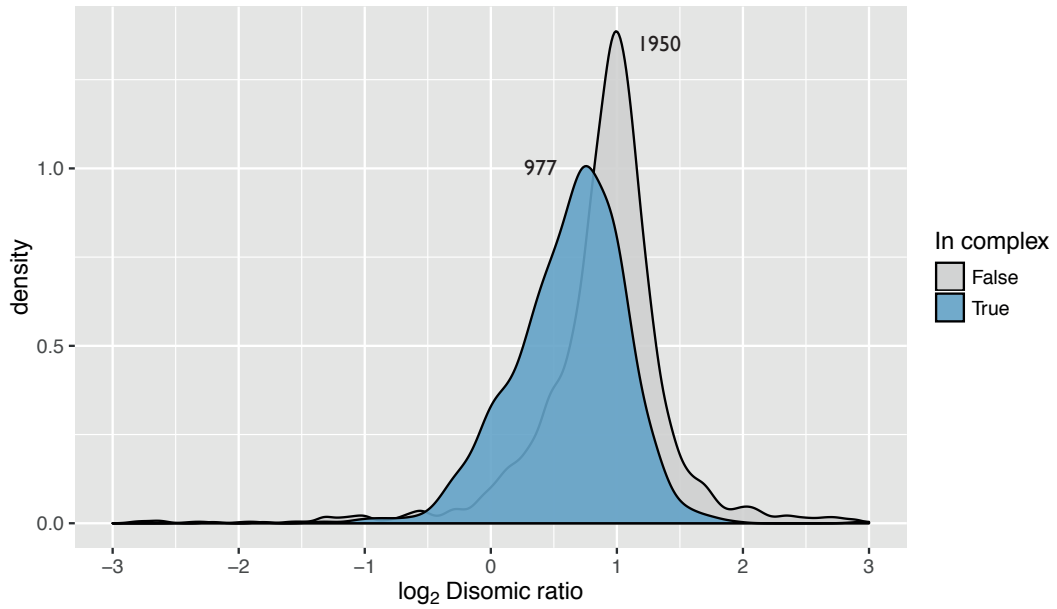


Figure A.11.: Replicate of fig. 5A-B, Dephoure et al.
Those proteins not found in the Pu et al.[307] dataset are approximately normally distributed around 1, i.e. are not attenuated at all upon gene duplication, with a median disomic ratio of 0.96. In contrast, proteins that are found in complexes are significantly attenuated, with a median of 0.68. 7 outliers with values < -3 or > 3 have been removed.

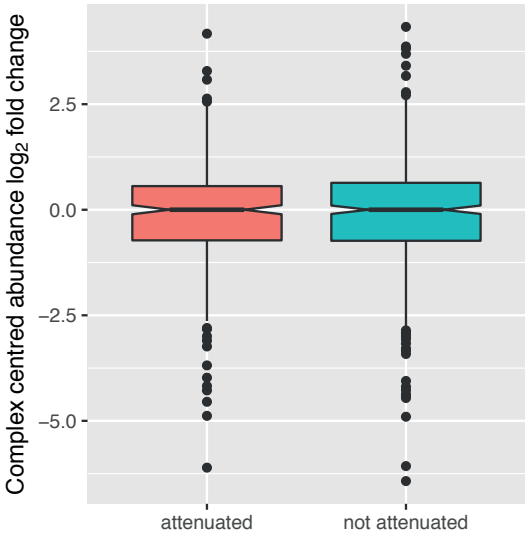


Figure A.12. Log2 fold-change in subunit abundance vs. median subunit abundance
When calculating the fold change of subunit abundance relative to the median subunit abundance within a complex, there is no significant difference between attenuated and non-attenuated proteins. This is in contrast to NED vs. ED, in which the former tend to be relatively more abundant. Fold change was calculated as log2 (subunit abundance/median abundance).

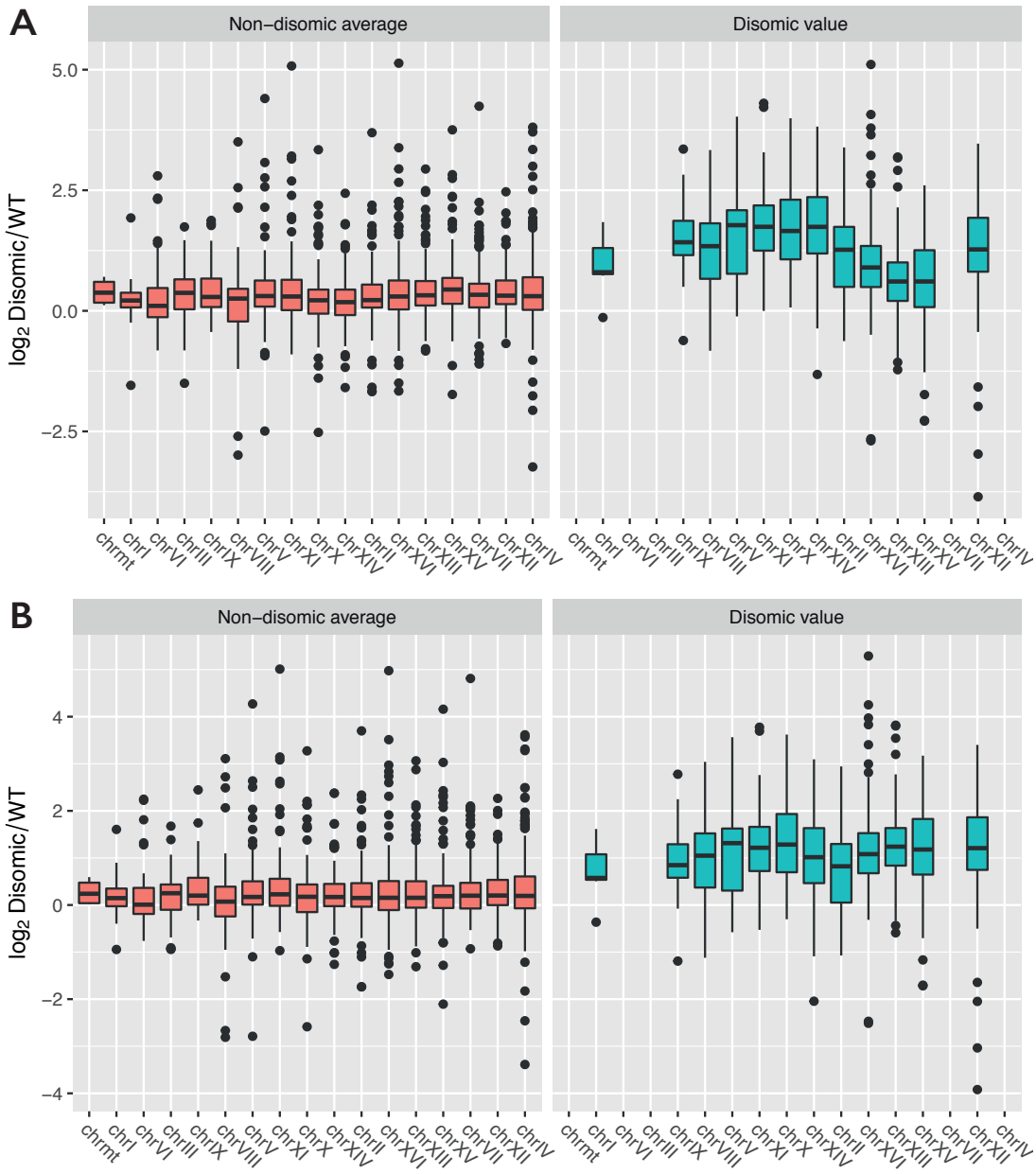


Figure A.13.: Pre- and post-normalisation of aggregation data
Pilot aggregation data pre- (A) and post-normalisation (B). Chromosomes are arranged ordered by size, from smallest to largest, as measured by number of genes. 'Non-disomic average' refers to all proteins on non-duplicated chromosomes, averaged across all experiments in which they are detected. Normalisation carried out by...

A.1.4 Chapter 5: Hawk proteins: A paralogous family of eukaryotic SMC-kleisin regulators

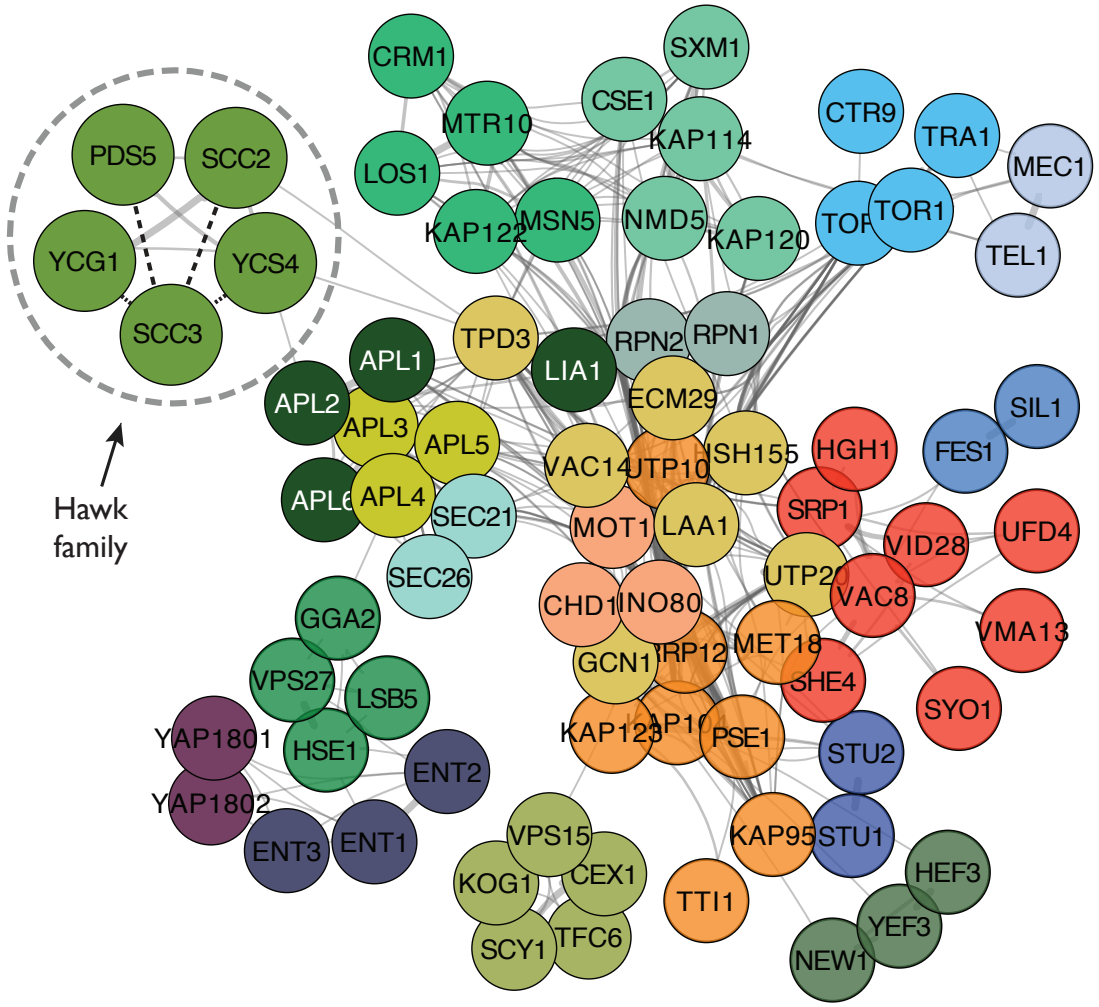


Figure A.14.: Clustered yeast network with inter-cluster edges
Saccharomyces cerevisiae network with inter-cluster edges retained. Only edges with HHsearch true positive probability greater than 99.5% are shown for the sake of clarity. Related to figure 5.3. Clathrin adaptors are shown with white labels. Adapted from figure 1, Wells et al.[3]

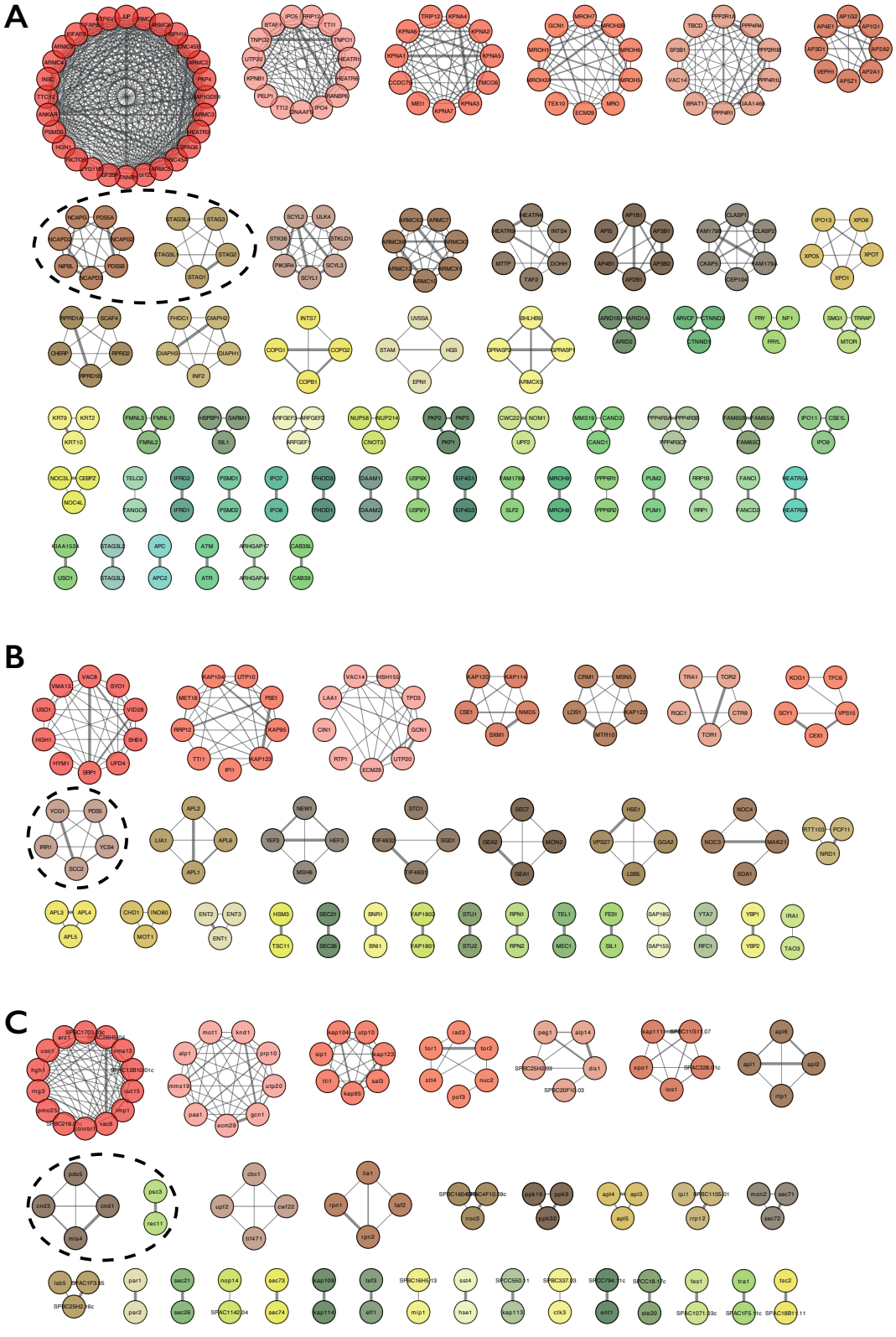


Figure A.15.: Homology networks from human and fission yeast
Clustered networks from *H. sapiens* (A), *S. cerevisiae* (B), and *S.pombe* (C), with inter-cluster edges removed. Hawk clusters are shown within dashed rings. Adapted from figure S1, Wells et al.[3]

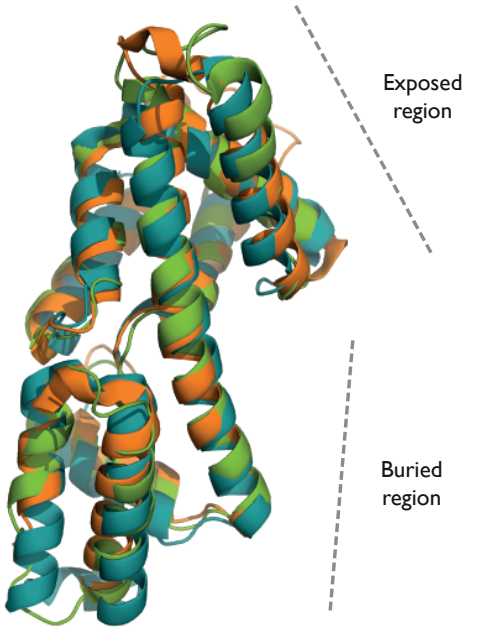


Figure A.16. Pds5 indel from three species
Structural alignment of the indel region from Pds5/B in H. sapiens (teal), S. cerevisiae (green) and L. thermatolerans (orange, 5HDT, 5FRR and 5F0N respectively, marked with asterisk in figure 5.4). Whilst there is no clear sequence conservation, the extended alpha-helix (centre) is apparently a defining feature of the region. Adapted from figure S2, Wells et al.[3]

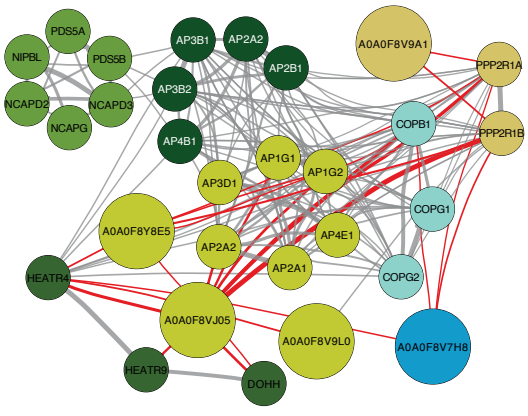


Figure A.17. Lokiarchaeal HEAT repeat proteins integrated into human network
Lokiarchaeal HEAT-like proteins (larger circles, red edges) show no indication of being directly related to the hawks, but do show highly significant similarity to one subgroup of the clathrin adaptor proteins, supporting recent evidence of an archaeal origin for these proteins[346]

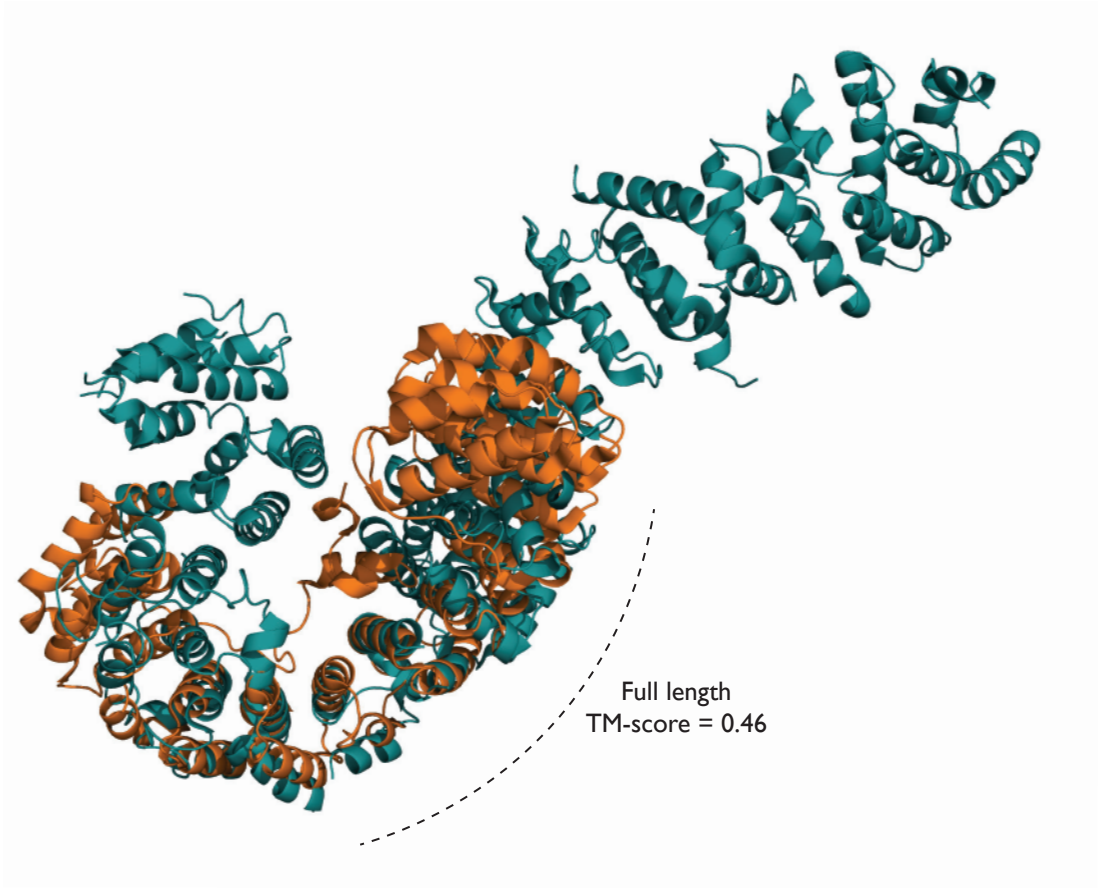


Figure A.18.: Structural similarity between hawks and clathrin adaptors
Structural alignment of *L. thermatolerans* Pds5 (5F0O, teal) and human AP2B (2XA7, orange). The TM-score of 0.46 indicates that the similarity is unlikely to be due to chance alone; having said that, care should be taken not to over-interpret the similarity as it is relatively to achieve good structural alignments of repeat proteins such as these.

H. sapiens		*S. cerevisiae*		*S.pombe*	
Gene	Cluster	Gene	Cluster	Gene	Cluster
RICTOR	1	SYO1	1	pmo25	1
HSF2BP	1	HYM1	1	SPBC216.01c	1
HGH1	1	HGH1	1	ctnnbl1	1
ZYG11B	1	USO1	1	rng3	1
PSMD5	1	VMA13	1	uso1	1
ANKAR	1	SHE4	1	hgh1	1
TTC12	1	VAC8	1	SPBC1703.03c	1
CTNNB1	1	VID28	1	arz1	1
INSC	1	UFD4	1	vma13	1
ARMC4	1	SRP1	1	SPAC12B10.01c	1
JUP	1	CIN1	2	SPAC26H5.04	1
ARMC9	1	RTP1	2	imp1	1
KIFAP3	1	LAA1	2	cut15	1
CFAP69	1	VAC14	2	vac8	1
UNC45B	1	HSH155	2	ecm29	2
ARMC8	1	TPD3	2	paa1	2
ATP6V1H	1	UTP20	2	mms19	2
ARMC2	1	ECM29	2	prp10	2
ARMC6	1	GCN1	2	alp1	2
RSPH14	1	IPI1	3	mot1	2
PKP4	1	TTI1	3	utp20	2
ARMC3	1	UTP10	3	knd1	2
RAP1GDS1	1	MET18	3	gcn1	2
HEATR3	1	RRP12	3	tti1	3
UNC45A	1	PSE1	3	utp10	3
SPAG6	1	KAP104	3	sip1	3
C1orf228	1	KAP123	3	kap104	3
ARMC5	1	KAP95	3	kap123	3
TTI2	2	SCY1	4	sal3	3
PELP1	2	KOG1	4	kap95	3
DNAAF5	2	TFC6	4	pof3	4
KPNB1	2	VPS15	4	nuc2	4
TNPO2	2	CEX1	4	rad3	4
UTP20	2	RQC1	5	str4	4
IPO5	2	TRA1	5	tor2	4
TNPO1	2	CTR9	5	tor1	4
TTI1	2	TOR2	5	SPBC20F10.03	5
BTAF1	2	TOR1	5	dis1	5
HEATR1	2	CSE1	6	peg1	5
HEATR6	2	KAP120	6	SPBC25H2.03	5
RANBP6	2	NMD5	6	alp14	5
RRP12	2	KAP114	6	xpo1	6
IPO4	2	SXM1	6	los1	6
MEI1	3	YCG1	7	kap111	6
CCDC79	3	YCS4	7	SPAC328.01c	6
KPNA7	3	PDS5	7	SPBC11G11.07	6
KPNA4	3	IRR1	7	upf2	7
TMCO6	3	SCC2	7	cbc1	7
KPNA5	3	LOS1	8	cwf22	7
KPNA6	3	MSN5	8	tif471	7
KPNA1	3	CRM1	8	rpn1	8
KPNA2	3	KAP122	8	rpn2	8
TRIP12	3	MTR10	8	taf2	8
KPNA3	3	SDA1	9	lia1	8
TEX10	4	NOC3	9	rtp1	9
MROH2A	4	NOC4	9	apl1	9
ECM29	4	MAK21	9	apl2	9
MROH1	4	TIF4632	10	apl6	9
MROH7	4	TIF4631	10	pds50	10
MROH6	4	STO1	10	cnd10	10
GCN1	4	SGD1	10	cnd30	10

H. sapiens		*S. cerevisiae*		*S.pombe*	
Gene	Cluster	Gene	Cluster	Gene	Cluster
MROH2B	4	GEA2	11	mis40	10
MROH5	4	GEA1	11	ppk321	11
MRO	4	SEC7	11	ppk191	11
BRAT1	5	MON2	11	ppk31	11
SF3B1	5	MSH6	12	mon22	12
PPP4R4	5	YEF3	12	sec712	12
VAC14	5	HEF3	12	sec722	12
TBCD	5	NEW1	12	SPAC4F10.09c3	13
PPP2R1A	5	HSE1	13	noc33	13
PPP2R1B	5	VPS27	13	SPBC1604.06c3	13
PPP4R1L	5	GGA2	13	lsb54	14
KIAA1468	5	LSB5	13	SPAC1F3.054	14
PPP4R1	5	LIA1	14	SPBC25H2.16c4	14
VEPH1	6	APL2	14	ipi15	15
AP5Z1	6	APL6	14	SPBC1105.015	15
AP1G1	6	APL1	14	rrp125	15
AP4E1	6	NRD1	15	apl46	16
AP2A1	6	RTT103	15	apl36	16
AP3D1	6	PCF11	15	apl56	16
AP2A2	6	CHD1	16	par27	17
AP1G2	6	INO80	16	par17	17
SCYL1	7	MOT1	16	SPAC18B11.118	18
PIK3R4	7	ENT2	17	tsc28	18
STK36	7	ENT3	17	SPBC337.039	19
SCYL2	7	ENT1	17	ctk39	19
SCYL3	7	APL3	18	mip10	20
ULK4	7	APL4	18	SPBC16H5.130	20
STKLD1	7	APL5	18	sec731	21
NCAPG	8	BNR1	19	sec741	21
NCAPD3	8	BNI1	19	hse12	22
NIPBL	8	YBP1	20	sst42	22
NCAPD2	8	YBP2	20	SPAC1142.043	23
PDS5B	8	HSM3	21	nop143	23
NCAPG2	8	TSC11	21	SPAC1071.03c4	24
PDS5A	8	SAP155	22	fes14	24
ARMC12	9	SAP185	22	rec115	25
ARMCX3	9	YAP1801	23	psc35	25
ARMCX6	9	YAP1802	23	tra16	26
ARMCX2	9	IRA1	24	SPAC1F5.11c6	26
ARMC7	9	TAO3	24	kap1137	27
ARMCX1	9	RPN1	25	SPCC550.117	27
ARMC10	9	RPN2	25	sec268	28
MTTP	10	SIL1	26	sec218	28
TAF2	10	FES1	26	ste209	29
HEATR9	10	YTA7	27	SPCC18.17c9	29
HEATR4	10	RFC1	27	kap1090	30
INTS4	10	TEL1	28	kap1140	30
DOHH	10	MEC1	28	ent11	31
API5	11	STU1	29	SPCC794.11c1	31
AP1B1	11	STU2	29	tef32	32
AP4B1	11	SEC21	30	elf12	32
AP2B1	11	SEC26	30		
AP3B2	11				
AP3B1	11				
CEP104	12				
CKAP5	12				
FAM179B	12				
CLASP1	12				
CLASP2	12				
FAM179A	12				
CHERP	13				

H. sapiens		*H. sapiens*	
Gene	Cluster	Gene	Cluster
RPRD1A	13	ARID1B	30
SCAF4	13	ARID2	30
RPRD2	13	ARID1A	30
RPRD1B	13	PKP1	31
STAG3L4	14	PKP2	31
STAG3	14	PKP3	31
STAG3L1	14	SIL1	32
STAG2	14	HSPBP1	32
STAG1	14	SARM1	32
DIAPH3	15	CTNND1	33
DIAPH2	15	ARVCF	33
FHDC1	15	CTNND2	33
DIAPH1	15	IPO11	34
INF2	15	CSE1L	34
XPO5	16	IPO9	34
IPO13	16	MMS19	35
XPO1	16	CAND2	35
XPO6	16	CAND1	35
XPOT	16	CAB39	36
EPN1	17	CAB39L	36
HGS	17	FANCI	37
STAM	17	FANCD2	37
UVSSA	17	FAM178B	38
INTS7	18	SLF2	38
COPB1	18	RRP1B	39
COPG1	18	RRP1	39
COPG2	18	PPP6R2	40
ARMCX5	19	PPP6R1	40
GPRASP2	19	USP9X	41
BHLHB9	19	USP9Y	41
GPRASP1	19	ARHGAP44	42
KRT2	20	ARHGAP17	42
KRT10	20	KIAA1524	43
KRT9	20	USO1	43
NOC4L	21	TELO2	44
NOC3L	21	TANGO6	44
CEBPZ	21	PUM1	45
ARFGEF2	22	PUM2	45
ARFGEF1	22	ATM	46
ARFGEF3	22	ATR	46
NUP58	23	PSMD2	47
CNOT3	23	PSMD1	47
NUP214	23	MROH9	48
CWC22	24	MROH8	48
UPF2	24	DAAM1	49
NOM1	24	DAAM2	49
TRRAP	25	EIF4G3	50
SMG1	25	EIF4G1	50
MTOR	25	FHOD1	51
FRYL	26	FHOD3	51
NF1	26	IFRD2	52
FRY	26	IFRD1	52
PPP4R3A	27	IPO8	53
PPP4R3B	27	IPO7	53
PPP4R3CP	27	STAG3L2	54
FMNL2	28	STAG3L3	54
FMNL1	28	HEATR5A	55
FMNL3	28	HEATR5B	55
FAM65C	29	APC	56
FAM65A	29	APC2	56
FAM65B	29		

Table A.1.: Complete list of hawk clusters

A.2 PUBLISHED PAPERS

This appendix includes all papers published during the course of my PhD studies. The first three are novel research publications that form the basis of the material in chapters 2-5 respectively, whereas the last two are reviews on the topic of co-translational assembly. These reviews are not used in this thesis other than as citations, and can be considered as supporting papers. In order of appearance, the papers included here are as follows:

1. Wells, J. N., Bergendahl, L. T. & Marsh, J. A. Operon Gene Order Is Optimized for Ordered Protein Complex Assembly. *Cell Reports* **14,** 679–685. ISSN: 22111247 (Feb. 2016)

2. McShane, E., Sin, C., Zauber, H., Wells, J. N., Donnelly, N., Wang, X., Hou, J., Chen, W., Storchova, Z., Marsh, J. A., Valleriani, A. & Selbach, M. Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. *Cell* **167,** 803–815. ISSN: 00928674 (Oct. 2016)

3. Wells, J. N., Gligoris, T. G., Nasmyth, K. A. & Marsh, J. A. Evolution of condensin and cohesin complexes driven by replacement of Kite by Hawk proteins. *Current Biology* **27,** R17–R18. ISSN: 09609822 (Jan. 2017)

4. Wells, J. N., Bergendahl, L. T. & Marsh, J. A. Co-translational assembly of protein complexes. *Biochemical Society Transactions* **43,** 1221–1226. ISSN: 0300-5127 (Dec. 2015)

5. Natan, E., Wells, J. N., Teichmann, S. A. & Marsh, J. A. Regulation, evolution and consequences of cotranslational protein complex assembly. *Current Opinion in Structural Biology* **42,** 90–97. ISSN: 0959-440X (Feb. 2017)

Cell Reports
Report

Operon Gene Order Is Optimized for Ordered Protein Complex Assembly

Jonathan N. Wells,[1] L. Therese Bergendahl,[1] and Joseph A. Marsh[1,*]
[1]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, United Kingdom
*Correspondence: joseph.marsh@igmm.ed.ac.uk
http://dx.doi.org/10.1016/j.celrep.2015.12.085

SUMMARY

The assembly of heteromeric protein complexes is an inherently stochastic process in which multiple genes are expressed separately into proteins, which must then somehow find each other within the cell. Here, we considered one of the ways by which prokaryotic organisms have attempted to maximize the efficiency of protein complex assembly: the organization of subunit-encoding genes into operons. Using structure-based assembly predictions, we show that operon gene order has been optimized to match the order in which protein subunits assemble. Exceptions to this are almost entirely highly expressed proteins for which assembly is less stochastic and for which precisely ordered translation offers less benefit. Overall, these results show that ordered protein complex assembly pathways are of significant biological importance and represent a major evolutionary constraint on operon gene organization.

INTRODUCTION

The assembly of proteins into complexes is integral to a wide range of biological processes. Although we now have extensive knowledge of the diverse quaternary structures formed by protein complexes (Goodsell and Olson, 2000; Janin et al., 2008; Marsh and Teichmann, 2015; Ahnert et al., 2015), much less is known about how they assemble and how assembly is regulated. In recent years, advances in electrospray mass spectrometry techniques have provided major new insights into in vitro assembly, allowing the assembly and disassembly pathways of protein complexes with diverse quaternary structure topologies to be elucidated in detail (Hernández and Robinson, 2007). In homomers, formed from the self-assembly of a single type of polypeptide chain, experimentally identified assembly intermediates often correspond to putative evolutionary precursors, so that the evolutionary history of a complex is reflected in its assembly pathway (Levy et al., 2008). Heteromers, formed from multiple distinct subunits, also tend to assemble and disassemble via ordered pathways that have a strong tendency to be evolutionarily conserved (Marsh et al., 2013). Although these experiments can be time-consuming, ordered assembly pathways can usually be predicted with very good accuracy from the known three-dimensional structure of a complex (Levy et al., 2008; Marsh et al., 2013). Given the many thousands of protein complex structures that are now available, this enables the study of assembly on a larger scale using computationally predicted assembly pathways.

Within the cell, assembly is much more complex and stochastic than in vitro, particularly in heteromers where multiple protein-coding genes must first be transcribed to mRNA and translated into protein, and those proteins must then find each other and assemble. Assembly is especially difficult for lowly expressed proteins, for which the stochastic variations in relative subunit concentrations are greater and the probability of interaction is lower (Kovács et al., 2009; Swain et al., 2002). How do cells cope with this? Does assembly within the cell follow similar ordered pathways as those observed in vitro and predicted computationally? Where does assembly occur within the cell? Has the regulation of gene expression been optimized for protein complex assembly order, as appears to be the case for the large multi-subunit bacterial flagella (Kalir et al., 2001)? Here we were able to address all of these questions by considering the relationship between protein complex assembly and gene organization in prokaryotic operons.

RESULTS

Operon-Encoding of Protein Complexes Is Likely to Enhance the Efficiency of Assembly

Many operons contain genes encoding different subunits of the same protein complex (Dandekar et al., 1998; Mushegian and Koonin, 1996) that can then be transcribed onto the same polycistronic mRNA. We first searched for heteromeric protein complexes of known structure from all prokaryotic organisms where at least two of the subunits are encoded by different genes from the same operon. In total, we identified 368 non-redundant pairs of subunits from the same heteromer encoded by different genes from the same operon (Figure 1A, left) from 70 different bacterial and archaeal species. This compares to 711 pairs encoded by different transcriptional units (*i.e.* translated from different mRNAs) from the same species (Figure 1A, right).

It has been suggested previously that a major advantage of operon-encoded complexes is their more efficient assembly because of smaller stochastic fluctuations in relative concentration than would occur if separate transcription steps were required for each subunit (Shieh et al., 2015; Sneppen et al.,

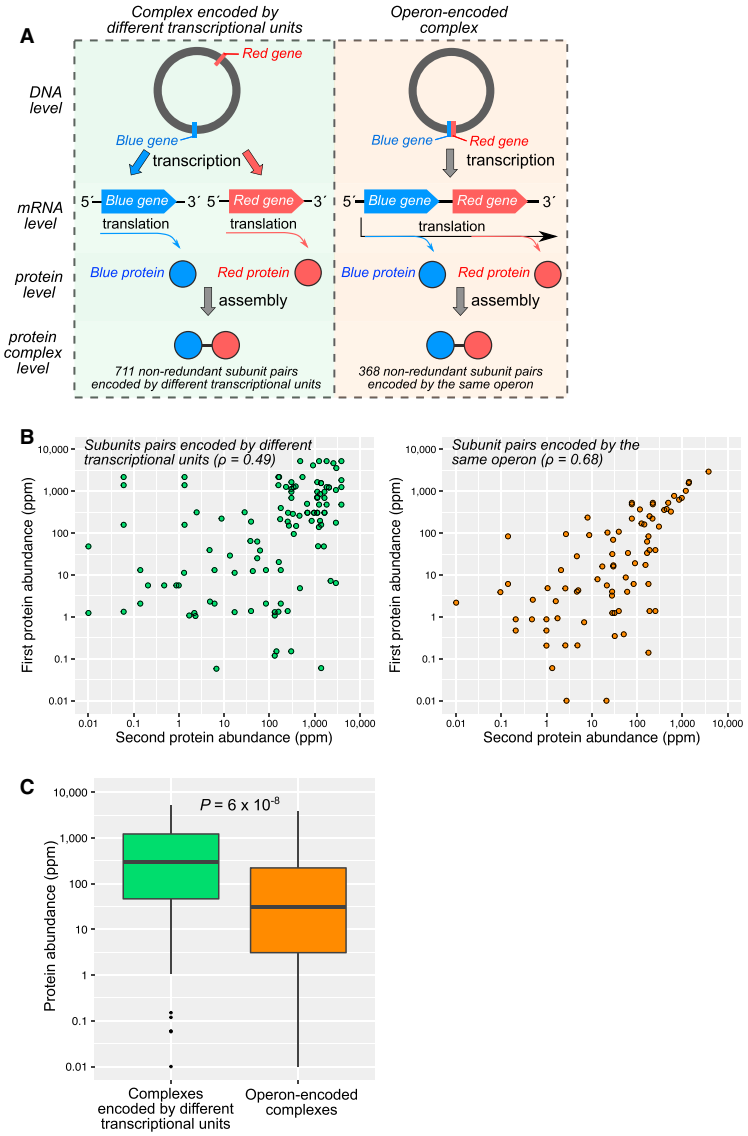


Figure 1. Operon Encoding of Protein Complex Subunits Enhances the Efficiency of Assembly

(A) Comparison of assembly for heterodimers where different subunits are encoded by different transcriptional units and where genes encoding both subunits are present on the same operon.

(B) Correlation (Spearman's ρ) between abundance measurements from subunit pairs encoded by different transcriptional units or by the same operon. The correlation for subunit pairs encoded by the same operon is significantly higher than for those encoded by different transcriptional units ($p = 0.002$), as calculated by randomly shuffling the pairs between two groups of the same size 10^5 times.

(C) Comparison of protein abundance measurements for subunits from operon-encoded complexes versus other subunits from complexes encoded by different transcriptional units. Boxes represent quartile distributions, and whiskers extend up to 1.5× the interquartile range. The p value was calculated with Wilcoxon rank-sum test.

Figure S1 shows these comparisons using protein abundance measurements combined from multiple organisms and with *E. coli* protein synthesis rates.

2010; Swain, 2004). Unsurprisingly, it has been observed that proteins encoded by the same operon tend to be coexpressed (Wang et al., 2005). Similarly, in Figure 1B, we demonstrate a stronger correlation in *E. coli* protein abundance measurements (Wang et al., 2015) between pairs of subunits encoded by the same operon compared with pairs of subunits encoded by different transcriptional units.

If operons do provide a mechanism to minimize the stochasticity of assembly, then we can further predict that lowly expressed complexes should be more likely to be in operons because their assembly is inherently more stochastic (Kovács et al., 2009; Swain et al., 2002). This is supported by a highly significant ($p = 6 \times 10^{-8}$) tendency for operon-encoded subunits to be lower in abundance than subunits from complexes encoded by different transcriptional units (Figure 1C). Although there is an overlap between the groups, this suggests that lowly expressed genes encoding interacting subunits may have experienced stronger evolutionary pressure to be located on the same operon because of their more stochastic assembly. Alternatively, because of the efficiency of their assembly, operon-encoded subunits may only need to be expressed at lower levels.

Operon-Encoded Subunits Tend to Be Encoded by Neighboring Genes and Form Large Interfaces

In addition to simply having genes encoding interacting subunits on the same operon, another way to enhance the efficiency of assembly would be to position the genes close together. If two genes encoding interacting subunits are close, then the newly translated subunits will also be close and more likely to encounter each other than if the two genes are farther apart (Figure 2A). In fact, the tendency for adjacent genes to code for interacting proteins has long been recognized (Dandekar et al., 1998; Mushegian and Koonin, 1996).

In Figure 2B, we plot the number of subunit pairs from the same complex by the distance between their genes within the operon. Strikingly, we see that 220 of 368 subunit pairs (59.8%) are

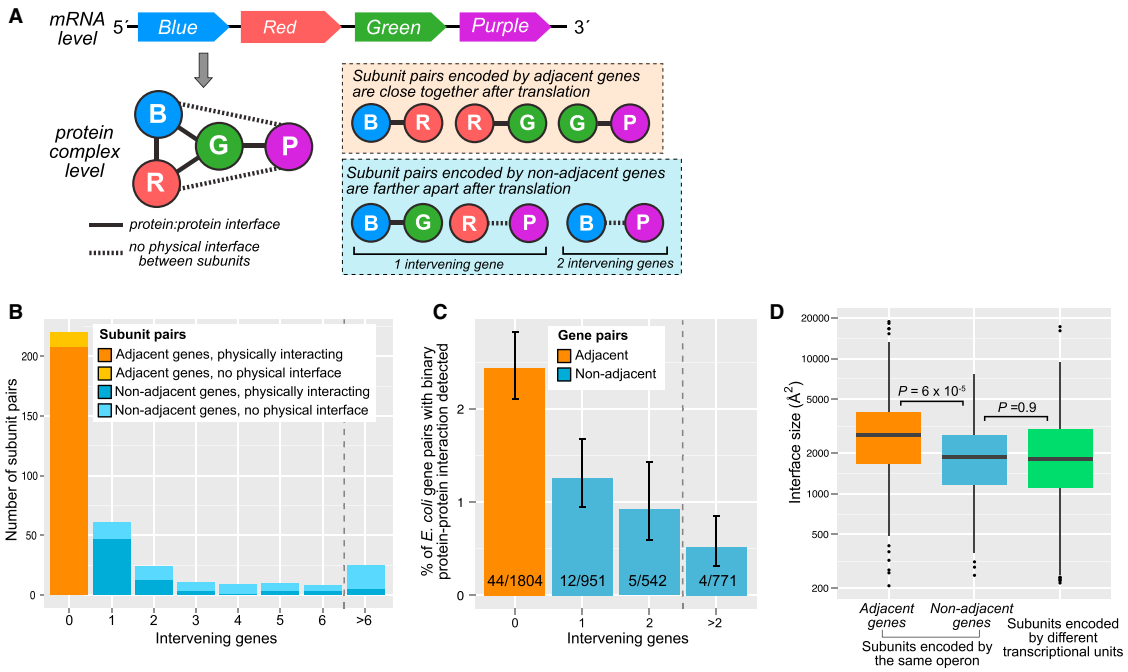


Figure 2. Genes Encoding Interacting Subunits of the Same Complex Tend to Be Close Together on an Operon

(A) Illustration of how operon structure can be related to quaternary structure with a hypothetical four-subunit heteromer. Pairs of subunits from the same complex can be encoded by genes that are adjacent on an operon or farther apart.

(B) Number of subunit pairs encoded by the same operon, grouped by the distance between their encoding genes. Subunit pairs are also divided into those that interact physically, which we define as forming an interface of >200 Å, and those that do not interact physically.

(C) Percentage of pairs of *E. coli* genes from the same operon for which a binary yeast two-hybrid interaction could be detected. Error bars represent 68% Wilson binomial confidence intervals.

(D) Distribution of interface sizes formed between physically interacting subunit pairs encoded by adjacent or non-adjacent genes on the same operon or between subunits encoded by different transcriptional units. Boxes represent quartile distributions, and whiskers extend up to 1.5× the interquartile range. The p values were calculated with Wilcoxon rank-sum test.

encoded by adjacent genes. Furthermore, because not all subunit pairs from the same complex physically interact with each other (e.g., blue-purple and red-purple in Figure 2A), we note that the tendency to form a physical intersubunit interface within the complex is much higher between the adjacent (208 of 220) compared with non-adjacent (77 of 148) pairs ($p = 5 \times 10^{-22}$, Fisher's exact test). Finally, this is supported further through analysis of a large set of *E. coli* binary protein-protein interactions (Rajagopala et al., 2014) where we confirmed that proteins encoded by adjacent genes are much more likely to interact (Figure 2C). Importantly, we show in Figure S2 that the tendency for interacting proteins to be close within an operon is highly significant compared with a null model in which gene order is randomized.

Figure 2D compares the sizes of interfaces formed between subunits encoded by adjacent genes, subunits encoded by non-adjacent genes from the same operon, and subunits encoded by different transcriptional units. We observe a highly significant tendency for adjacent subunits to be larger, although the interface size distribution is very broad and there is considerable overlap between the groups. This is especially interesting when considering that larger interfaces within a complex will usually assemble earlier than smaller interfaces (Levy et al., 2008; Marsh et al., 2013). This provides further evidence that operon structure appears to have been evolutionarily optimized for protein complex formation. Even when we consider only physically interacting proteins, those that form larger interfaces and are, therefore, likely to assemble earlier are much more likely to be encoded by adjacent genes.

The above observation could potentially have implications for our previous finding that evolutionary gene fusion events tend to conserve existing assembly pathways (Marsh et al., 2013) because fusion often occurs between adjacent genes. However, we show in Figure S3 that, even if only subunit pairs encoded by adjacent genes are considered, there still appears to be evolutionary selection for assembly-conserving fusions.

Operon Gene Order Is Optimized for the Order of Protein Complex Assembly

The above results suggest that operon-encoded subunits will often be synthesized very close to each other within the cell.

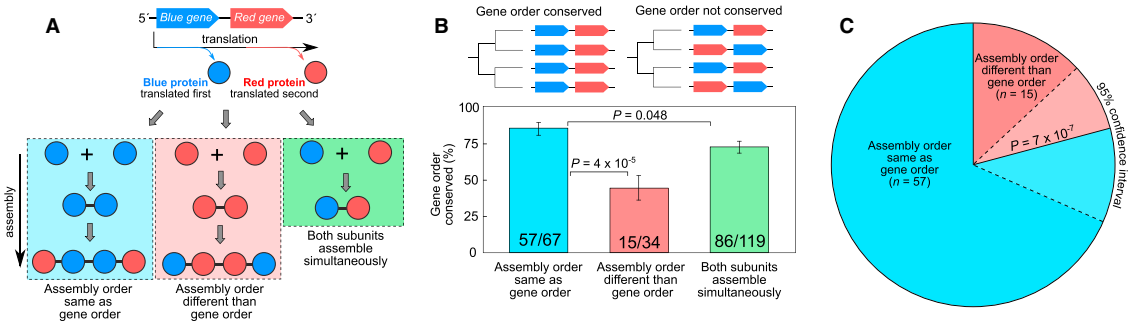


Figure 3. Operon Gene Order Reflects the Order of Protein Complex Assembly

(A) Illustration of the three possible relationships between gene pair order and subunit assembly order.

(B) Evolutionary conservation in pairs of adjacent genes encoding subunits of the same complex. The p values were calculated with Fisher's exact test. Error bars represent 68% Wilson binomial confidence intervals.

(C) When considering adjacent gene pairs with evolutionarily conserved gene order that encode different subunits of the same protein complex, the predicted assembly order is the same as the gene order in 57 of 72 cases. The p value was calculated with a binomial test.

However, there is also a temporal component to this in that upstream genes will tend to be translated before downstream genes. This is first due to coupled transcription and translation, where the upstream gene that is transcribed first will also be translated first (Gowrishankar and Harinarayanan, 2004), and second to translational coupling, in which translating ribosomes can continue on to downstream genes (Oppenheim and Yanofsky, 1980). Therefore, if genes are arranged so that the gene order matches the order of subunit assembly, then the newly translated subunits will be more likely to interact quickly.

We illustrate this in Figure 3A with the example of a hypothetical operon containing two adjacent genes, *blue* and *red*. If these genes encode different subunits of the same complex, then there are three possible relationships between gene order and assembly order. First, the assembly order could be the same as the gene order if the blue subunit that is translated first also assembles first. Second, the assembly order could be different than the gene order if the blue subunit assembles last. Finally, both subunits could assemble simultaneously, as would be the case for a simple heterodimer where the first step of assembly is the heteromeric interaction between different subunits.

Using our previous observation that assembly pathways can be predicted using interface sizes from three-dimensional structures of protein complexes (Marsh et al., 2013), here we predicted the assembly pathways for all operon-encoded heteromers in our dataset and classified each of the 220 adjacent gene pairs into one of these three groups. We then considered the tendency for gene order to be evolutionarily conserved in each group (Figure 3B). Interestingly, the evolutionary conservation of gene order is significantly higher in cases where it is the same as the predicted assembly order. This suggests that the evolutionary constraint on gene order is much stronger when it is optimized for assembly.

Next, we consider 72 gene pairs where gene order is evolutionarily conserved and where one subunit is predicted to assemble before the other. Figure 3C illustrates the striking correspondence between gene order and assembly order, with 57 pairs (79.2%) having the same assembly order as gene order ($p = 7 \times 10^{-7}$, binomial test). In contrast, when the gene order is not evolutionary conserved, only 10 of 29 gene pairs show correspondence between gene order and assembly order. Therefore, selection for ordered protein complex assembly appears to be a major evolutionary determinant of operon gene order.

We can also consider the relationship between gene order and assembly order for non-adjacent genes. Although the dataset is smaller, the relationship between gene order and assembly order appears to get weaker between genes that are more distant (Figure S4). This is likely due to weaker spatial and temporal coupling between non-adjacent genes that are translated farther apart from each other, as evidenced by the fact that subunits encoded by non-adjacent genes are much less likely to physically interact with each other (Figure 2B). Interestingly, the relationship between gene order and assembly order is stronger for proteins that interact physically, particularly those that form large interfaces. Similarly, subunit pairs encoded by adjacent genes where gene order and assembly order are the same tend to have significantly larger interfaces (Figure S4).

A possible alternative explanation for the correspondence between gene order and assembly order could be if earlier-assembling subunits need to be expressed at higher levels. Specifically, there is evidence of a linear relationship between expression levels and the proximity of genes to the start of operons (Lim et al., 2011; Nishizaki et al., 2007). This is weakly supported in the dataset used here, with proteins encoded by upstream genes showing a slight but not significant tendency to be more abundant (Table S1). Importantly, we find that protein expression levels show essentially no relationship with assembly and that gene order is a significantly better predictor of assembly order (Figure S5).

Operon Gene Order Is Most Important for the Assembly of Lowly Expressed Proteins

Despite the strong correspondence between protein complex assembly and operon organization, there is still discordance

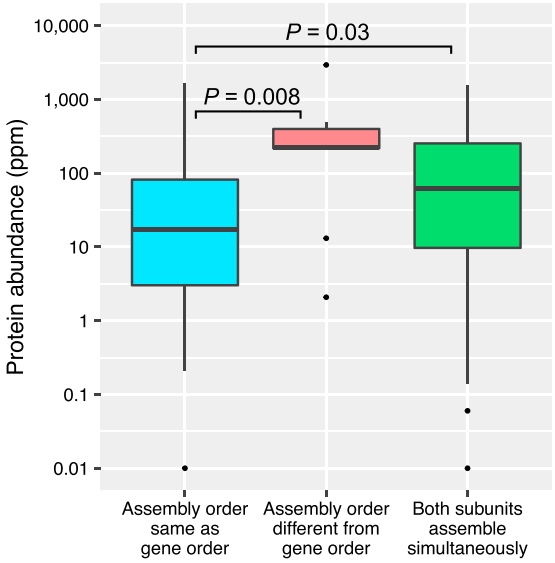


Figure 4. Cases Where Evolutionarily Conserved Gene Order Does Not Follow Assembly Order Tend to be Highly Expressed

Boxes represent quartile distributions of protein abundance measurements, and whiskers extend up to 1.5× the interquartile range. The p values were calculated with a Wilcoxon rank-sum test.

Figure S7 shows these comparisons using protein abundance measurements combined from multiple organisms and with *E. coli* protein synthesis rates.

between gene order and assembly order in >20% of cases where gene order is evolutionarily conserved. This suggests that there must be other factors besides assembly order that influence gene order conservation. For example, the operon order of enzyme genes is known to correlate with metabolic pathway order (Kovács et al., 2009; Zaslaver et al., 2006), although this seems unlikely to explain gene order in operon-encoded complexes. A search for gene ontology terms (Huntley et al., 2015) enriched in subunit pairs where gene order is either the same or different than assembly order revealed little that could account for the results observed here (Figure S6). Furthermore, if gene position can affect expression levels, as mentioned above, then there may be some evolutionary pressure to conserve gene order; for example, to not disrupt the relative subunit stoichiometry (Marsh et al., 2015).

The fact that operon gene order closely follows assembly order suggests that assembly must occur very shortly after protein synthesis because the more time newly synthesized subunits have to diffuse before assembly the less the order of gene expression should matter. Building on this, we hypothesize that the relationship between operon order and assembly order should be stronger for lowly expressed proteins. If they do not assemble quickly, diffusion will reduce the probability of two low-concentration subunits encountering each other. In contrast, the chance of interaction between highly expressed, abundant proteins will be greater, and so there is less need for assembly to occur close to the site of protein synthesis.

In Figure 4, we plot the distributions of intracellular protein abundance measurements for subunits where conserved gene order follows assembly and for those where it does not. Those proteins where assembly order is the same as gene order tend to be much lower in abundance ($p = 0.008$, Wilcoxon test). Therefore, it appears that the correspondence between gene order and assembly order can mostly be attributed to lowly expressed proteins for which assembly is more stochastic. Interestingly, subunits where both assemble simultaneously are intermediate in abundance, consistent with the fact that gene order should show no correspondence with assembly in these cases.

DISCUSSION

Overall, a number of important conclusions can be drawn from these results. First, protein complex assembly within the cell appears to often follow the same ordered pathways that can be characterized experimentally and predicted computationally, at least in the case of operon-encoded complexes. Although there will certainly be some exceptions, particularly in cases where assembly chaperones are involved or subunits are translated in different parts of the cell, these results strongly support the physiological relevance of using in vitro or computational methods to study assembly.

Second, the remarkable correspondence between predicted assembly order and gene order further validates the utility of structure-based assembly predictions. Given the huge number of protein complex structures now known, this opens the door to future large-scale analyses of protein assembly pathways and their regulation, evolution, and role in biological function and disease.

This work also tells us something about where assembly occurs within the cell. For the low-abundance, operon-encoded complexes studied here, assembly must occur very close to the site of translation for gene order to have such a significant effect. In some cases, assembly may even occur co-translationally, involving at least one nascent chain still in the process of being translated (Duncan and Mata, 2011; Wells et al., 2015), as has been demonstrated recently for the operon-encoded bacterial luciferase complex (Shieh et al., 2015).

Finally, these results strongly support the biological importance of assembly pathways and suggest that co-ordinating both the timing and location of translation is important for maximizing the efficiency of stochastic protein complex assembly. The fact that operon gene order has been optimized for assembly order in many protein complexes suggests that assembly order is often very important and that there is significant benefit from tightly co-ordinating gene expression and protein assembly. Given that eukaryotes do not have operons that allow multiple protein subunits to be translated from the same polycistronic mRNA, it will be interesting to systematically investigate which other mechanisms might be employed to enhance the efficiency of assembly.

EXPERIMENTAL PROCEDURES

Protein Structural Datasets

We started with the full set of prokaryotic X-ray and electron microscopy structures in the PDB on June 12, 2014. We considered all heteromeric pairs of

subunits from the same complex, defined as having at least two different protein chains of ≥30 residues each and mapping to different UniProt sequences from a single species. Complexes with known quaternary structure assignment errors (Levy, 2007) were excluded. Very large complexes with >24 subunits were excluded, because we have not shown that the assembly of these can be predicted accurately from their structures. Heteromeric subunit pairs were filtered for redundancy at the level of 50% sequence identity.

Mapping Subunit Pairs to Operons

Operon datasets were downloaded from the DOOR^2 database (Mao et al., 2014). Relevant datasets were identified based on the species and strain of each gene pair. After converting GI numbers to UniProt accession identifiers in each dataset, the set of gene pairs was mapped to the operon data. Operons encoding both members of a pair were added to a reference dictionary, with the locus and directionality of each gene being used to arrange constituent genes in order of expression. In rare cases where the copy number of a gene within an operon was found to be greater than one, the position of the gene in the operon was taken to be that of the first copy to be encountered, reading in the 5′ to 3′ direction. The set was then filtered to remove redundant operons (i.e., identical operons from similar strains or species). In total, 368 gene pairs (220 adjacent) were mapped to 192 unique operons, with the remaining 711 pairs being expressed in different transcriptional units. These are provided in Dataset S1. Similarly, we also mapped a set of 2,562 binary protein-protein interactions (IM-22059) (Rajagopala et al., 2014) to the *E. coli* K-12 W3110 operons to calculate the result in Figure 2C (provided in Dataset S2).

To assess whether the gene order of a pair was evolutionary conserved, we used the STRING v9.1 database (Franceschini et al., 2013). For each pair, we manually assessed, using the STRING online interface, whether all occurrences of a given gene pair shared the same gene order within their local evolutionary group as defined in STRING. This is at the level of phylum (*e.g.* Firmicutes or Euryarchaeota) or class for proteobacteria, with all groups provided in Dataset S1. Gene pairs present across only a very limited evolutionary range (less than three genera) were not considered to be evolutionarily conserved. Gene pairs associated with evolutionary gene fusion events were identified as those sharing >40% sequence identity with a gene pair with evidence for fusion in STRING, similar to what has been done previously (Marsh et al., 2013).

Abundance Measurements

We mapped all protein complex subunits in our dataset against the sequences of prokaryotic proteins from PaxDB v4.0 (Wang et al., 2015), selecting abundance measurements with >90% sequence identity to a subunit. The results in Figures 1 and 4 only use abundance measurements from *E. coli*, but the analyses in the Figures S1, S5, and S7 and Table S1 are repeated using combined measurements from all available prokaryotes and also using protein synthesis rates derived from ribosomal profiling (Li et al., 2014).

Prediction of Assembly Pathways

Ordered protein complex assembly pathways were predicted in a manner very similar to what has been done previously (Marsh et al., 2013). First, the complex is considered in terms of its constituent subunits and the sizes of the interfaces that can be formed between any pair of subunits are calculated with AREAIMOL (Winn et al., 2011). Our model assumes that assembly will proceed via formation of the largest possible interface. The process is then repeated by calculating all possible interfaces that could form between subunits and subcomplexes until the full complex is assembled. To define which of a pair of subunits assembles first and which assembles later, we consider the first step of assembly that brings the two subunits together within the same (sub)complex. Whichever subunit was part of a larger subcomplex prior to this step is defined as assembling first. For example, in the blue pathway in Figure 3A, the blue subunit homodimerizes first and then interacts sequentially with the free red subunits, so the blue subunit is defined as assembling first. If, alternatively, the first step of assembly had been a heterodimerization between the blue and red subunits, then both subunits would be classified as assembling simultaneously. The relative order of assembly for each subunit pair is included in Dataset S1, and all predicted assembly pathways are provided in Dataset S3. The source code for predicting assembly pathways from protein complex structures is available at http://github.com/marshlab/assembly-prediction.

The full set of gene ontology associations for complexes where assembly order and gene order are the same or different is provided in Dataset S4.

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures, one table, and four datasets and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2015.12.085.

AUTHOR CONTRIBUTIONS

J.M. conceived and designed the research. J.W., T.B., and J.M. performed the computational analyses. J.M. wrote the manuscript with contributions from all authors.

ACKNOWLEDGMENTS

We thank Sarah Teichmann for helpful discussions and comments on the manuscript. This work was supported by a University of Edinburgh Chancellor's Fellowship and Medical Research Council Career Development Award MR/M02122X/1 (to J.A.M.).

Received: August 25, 2015
Revised: November 7, 2015
Accepted: December 17, 2015
Published: January 21, 2016

REFERENCES

Ahnert, S.E., Marsh, J.A., Hernández, H., Robinson, C.V., and Teichmann, S.A. (2015). Principles of assembly reveal a periodic table of protein complexes. Science *350*, aaa2245.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. *23*, 324–328.

Duncan, C.D.S., and Mata, J. (2011). Widespread cotranslational formation of protein complexes. PLoS Genet. *7*, e1002398.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L.J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. *41*, D808–D815.

Goodsell, D.S., and Olson, A.J. (2000). Structural symmetry and protein function. Annu. Rev. Biophys. Biomol. Struct. *29*, 105–153.

Gowrishankar, J., and Harinarayanan, R. (2004). Why is transcription coupled to translation in bacteria? Mol. Microbiol. *54*, 598–603.

Hernández, H., and Robinson, C.V. (2007). Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. Nat. Protoc. *2*, 715–726.

Huntley, R.P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M.J., and O'Donovan, C. (2015). The GOA database: gene Ontology annotation updates for 2015. Nucleic Acids Res. *43*, D1057–D1063.

Janin, J., Bahadur, R.P., and Chakrabarti, P. (2008). Protein-protein interaction and quaternary structure. Q. Rev. Biophys. *41*, 133–180.

Kalir, S., McClure, J., Pabbaraju, K., Southward, C., Ronen, M., Leibler, S., Surette, M.G., and Alon, U. (2001). Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. Science *292*, 2080–2083.

Kovács, K., Hurst, L.D., and Papp, B. (2009). Stochasticity in protein levels drives colinearity of gene order in metabolic operons of Escherichia coli. PLoS Biol. *7*, e1000115.

Levy, E.D. (2007). PiQSi: protein quaternary structure investigation. Structure *15*, 1364–1367.

Levy, E.D., Boeri Erba, E., Robinson, C.V., and Teichmann, S.A. (2008). Assembly reflects evolution of protein complexes. Nature *453*, 1262–1265.

Li, G.-W., Burkhardt, D., Gross, C., and Weissman, J.S. (2014). Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell *157*, 624–635.

Lim, H.N., Lee, Y., and Hussein, R. (2011). Fundamental relationship between operon organization and gene expression. Proc. Natl. Acad. Sci. USA *108*, 10626–10631.

Mao, X., Ma, Q., Zhou, C., Chen, X., Zhang, H., Yang, J., Mao, F., Lai, W., and Xu, Y. (2014). DOOR 2.0: presenting operons and their functions through dynamic and integrated views. Nucleic Acids Res. *42*, D654–D659.

Marsh, J.A., and Teichmann, S.A. (2015). Structure, dynamics, assembly, and evolution of protein complexes. Annu. Rev. Biochem. *84*, 551–575.

Marsh, J.A., Hernández, H., Hall, Z., Ahnert, S.E., Perica, T., Robinson, C.V., and Teichmann, S.A. (2013). Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell *153*, 461–470.

Marsh, J.A., Rees, H.A., Ahnert, S.E., and Teichmann, S.A. (2015). Structural and evolutionary versatility in protein complexes with uneven stoichiometry. Nat. Commun. *6*, 6394.

Mushegian, A.R., and Koonin, E.V. (1996). Gene order is not conserved in bacterial evolution. Trends Genet. *12*, 289–290.

Nishizaki, T., Tsuge, K., Itaya, M., Doi, N., and Yanagawa, H. (2007). Metabolic engineering of carotenoid biosynthesis in Escherichia coli by ordered gene assembly in Bacillus subtilis. Appl. Environ. Microbiol. *73*, 1355–1361.

Oppenheim, D.S., and Yanofsky, C. (1980). Translational coupling during expression of the tryptophan operon of Escherichia coli. Genetics *95*, 785–795.

Rajagopala, S.V., Sikorski, P., Kumar, A., Mosca, R., Vlasblom, J., Arnold, R., Franca-Koh, J., Pakala, S.B., Phanse, S., Ceol, A., et al. (2014). The binary protein-protein interaction landscape of Escherichia coli. Nat. Biotechnol. *32*, 285–290.

Shieh, Y.-W., Minguez, P., Bork, P., Auburger, J.J., Guilbride, D.L., Kramer, G., and Bukau, B. (2015). Operon structure and cotranslational subunit association direct protein assembly in bacteria. Science *350*, 678–680.

Sneppen, K., Pedersen, S., Krishna, S., Dodd, I., and Semsey, S. (2010). Economy of operon formation: cotranscription minimizes shortfall in protein complexes. MBio *1*, e00177–e10.

Swain, P.S. (2004). Efficient attenuation of stochasticity in gene expression through post-transcriptional control. J. Mol. Biol. *344*, 965–976.

Swain, P.S., Elowitz, M.B., and Siggia, E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc. Natl. Acad. Sci. USA *99*, 12795–12800.

Wang, R., Prince, J.T., and Marcotte, E.M. (2005). Mass spectrometry of the M. smegmatis proteome: protein expression levels correlate with function, operons, and codon bias. Genome Res. *15*, 1118–1126.

Wang, M., Herrmann, C.J., Simonovic, M., Szklarczyk, D., and von Mering, C. (2015). Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics *15*, 3163–3168.

Wells, J.N., Bergendahl, L.T., and Marsh, J.A. (2015). Co-translational assembly of protein complexes. Biochem. Soc. Trans. *43*, 1221–1226.

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., et al. (2011). Overview of the *CCP*4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. *67*, 235–242.

Zaslaver, A., Mayo, A., Ronen, M., and Alon, U. (2006). Optimal gene partition into operons correlates with gene functional order. Phys. Biol. *3*, 183–189.

Article

Cell

Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation

Erik McShane,[1] Celine Sin,[2] Henrik Zauber,[1] Jonathan N. Wells,[3] Neysan Donnelly,[4] Xi Wang,[1] Jingyi Hou,[1] Wei Chen,[1,7] Zuzana Storchova,[4,5] Joseph A. Marsh,[3] Angelo Valleriani,[2] and Matthias Selbach[1,6,8,*]

[1]Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str.10, 13092 Berlin, Germany
[2]Department of Theory and Bio-Systems, Max Planck Institute of Colloids and Interfaces, 14424 Potsdam, Germany
[3]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, Western General Hospital, University of Edinburgh, Edinburgh EH4 3BB, UK
[4]Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany
[5]University of Kaiserslautern, Paul-Ehrlich Str. 24, 67663 Kaiserslautern, Germany
[6]Charité-Universitätsmedizin Berlin, 10117 Berlin, Germany
[7]Present address: Department of Biology, South University of Science and Technology of China, 1088 Xueyuan Road, Xili, Nanshan District, Shenzhen, Guangdong 518055, China
[8]Lead Contact
*Correspondence: matthias.selbach@mdc-berlin.de
http://dx.doi.org/10.1016/j.cell.2016.09.015

SUMMARY

Do young and old protein molecules have the same probability to be degraded? We addressed this question using metabolic pulse-chase labeling and quantitative mass spectrometry to obtain degradation profiles for thousands of proteins. We find that >10% of proteins are degraded non-exponentially. Specifically, proteins are less stable in the first few hours of their life and stabilize with age. Degradation profiles are conserved and similar in two cell types. Many non-exponentially degraded (NED) proteins are subunits of complexes that are produced in super-stoichiometric amounts relative to their exponentially degraded (ED) counterparts. Within complexes, NED proteins have larger interaction interfaces and assemble earlier than ED subunits. Amplifying genes encoding NED proteins increases their initial degradation. Consistently, decay profiles can predict protein level attenuation in aneuploid cells. Together, our data show that non-exponential degradation is common, conserved, and has important consequences for complex formation and regulation of protein abundance.

INTRODUCTION

Pioneering experiments by Rudolph Schönheimer established that proteins are in a dynamic state of synthesis and degradation (Schoenheimer, 1942). The subsequent discovery of lysosomes and the ubiquitin-proteasome system (UPS) provided detailed insights into the molecular mechanisms of cellular protein homeostasis. It is now well-established that proteins are extensively turned over, that this process is specific, and that the stability of individual proteins can vary under different physiological conditions (Ciechanover, 2005).

Despite these mechanistic insights, the kinetics of cellular protein degradation are still not well understood. Early analyses indicated that intracellular protein degradation follows first order kinetics (Goldberg and Dice, 1974; Schimke and Doyle, 1970). Accordingly, protein degradation is thought to be an exponential decay process in which young and old proteins have the same degradation probability per unit time (i.e., degradation rate) (Figure 1A). However, there is substantial evidence that protein degradation does not always follow first-order kinetics. Pulse-chase experiments by Wheatley et al. (1980) indicated that a substantial fraction of proteins are degraded within the first 2 hr after synthesis. In addition, the newly synthesized immature forms of proteins, like the cystic fibrosis transmembrane conductance regulator (CFTR) and basigin (CD147), were found to be rapidly degraded while the "older" mature forms were stable (Tyler et al., 2012; Ward and Kopito, 1994). It was also observed that proteins can be ubiquitinated co-translationally (Duttler et al., 2013), that most ubiquitinated proteins in a cell are relatively young (Kim et al., 2011), and that MHC class I peptides sometimes show higher turnover than their source proteins (Bourdetsky et al., 2014). Finally, it has been estimated that ~30% of newly synthesized proteins are quickly degraded (Schubert et al., 2000), although this number was questioned in later studies (Vabulas and Hartl, 2005). Collectively, these studies suggest that decay probabilities of proteins can vary as a function of their molecular age. However, to the best of our knowledge, the degradation kinetics of individual proteins has not yet been investigated globally.

To systematically assess cellular protein degradation kinetics, we sought to perform pulse-chase experiments on a proteome-wide scale. The general idea is to metabolically label a population of proteins with a short pulse and then to quantify how much of this population is left after different lengths of chase. We and others have previously used stable isotope labeling by amino acids in cell culture (SILAC) to study protein synthesis and turnover (Andersen et al., 2005; Doherty et al., 2009; Hinkson and Elias, 2011; Jovanovic et al., 2015; Kristensen et al., 2013; Larance et al., 2013; Schwanhäusser et al., 2011; Selbach et al., 2008). A disadvantage of these approaches is that they

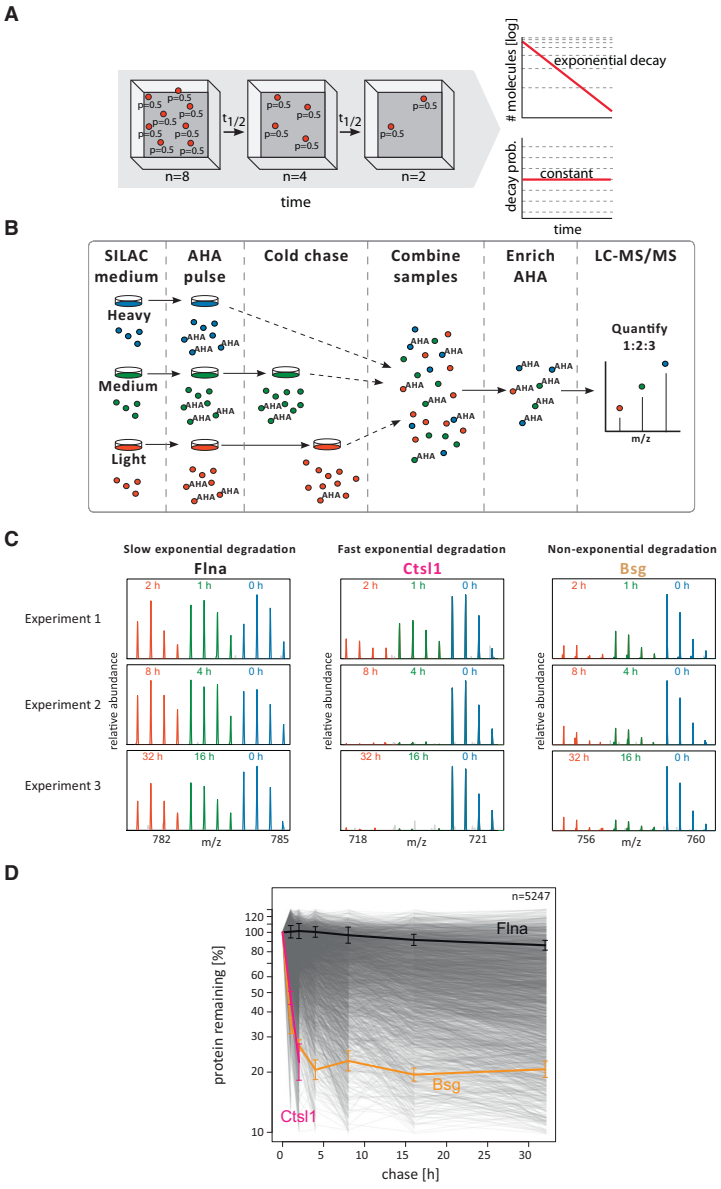


Figure 1. Global Quantification of Protein Degradation Kinetics by AHA Pulse-Chase

(A) Exponential decay can be recognized as a straight line (in a semi log plot) indicating that the degradation rate is constant, i.e., young and old molecules have the same degradation probability per unit time.

(B) Experimental setup for global pulse-chase experiments. SILAC-labeled cells are pulse-labeled with azidohomoalanine (AHA) and either directly harvested (time point 0 hr) or chased in medium without AHA (cold chase). Samples are combined, AHA containing proteins are enriched, digested, and analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS).

(C) Measured MS1 spectra for three peptides representing the major types of decay profiles detected. Filamin alpha (Flna, ASGPGLNTTGVPAS LPVEFTIDAK) shows slow exponential degradation, cathepsin L1 (Ctsl1, NLDHGVLLVGYGYEGTDSNK) shows fast exponential degradation, basigin (Bsg, VLQEDTLPDLHTK) shows non-exponential degradation.

(D) Decay profiles of individual proteins based on the median of three biological replicates (gray traces). Note that due to the experimental design not all proteins were detected at all time points. Increases in protein levels over time are theoretically impossible and probably reflect measurement noise. Highlighted profiles depict proteins shown in (C) and are based on all three replicates (mean ± SD). Outliers (>130% protein left) were removed.

See also Figures S1, S2, and S3.

require rather long labeling times. Metabolic labeling with bioorthogonal amino acids has emerged as an attractive alternative (Dieterich et al., 2006). Cells can incorporate the artificial amino acid azidohomoalanine (AHA) into newly synthesized proteins instead of methionine (Kiick et al., 2002). AHA contains an azide group enabling capture of proteins via click chemistry (Dieterich et al., 2006). Combining AHA with SILAC enables relatively short pulse times (Eichelbaum and Krijgsveld, 2014; Eichelbaum et al., 2012).

Here, we combined AHA and SILAC to obtain a global survey of protein degradation kinetics. We find that a sizable fraction of proteins are degraded non-exponentially. Many non-exponentially degraded (NED) proteins are members of heteromeric protein complexes that are over-produced relative to other members of the same complex. Thus, in contrast to recent findings in bacteria (Li et al., 2014), disproportional protein synthesis appears to be common and evolutionarily conserved in metazoans. Our data allowed us to predict how protein levels change in response to gene copy-number alterations in aneuploid cells. Global quantification of protein degradation kinetics reveals an unexpected layer of posttranslational regulation with important functional implications.

RESULTS

Combining Metabolic Pulse Labeling and Click-Chemistry for Global Pulse-Chase Experiments

To perform proteome-wide pulse-chase experiments we combined AHA and SILAC labeling (Figure 1B). First, cells are fully

labeled heavy, medium-heavy, or light using SILAC. Second, all three cell populations are pulse-labeled with AHA for 1 hr. This relatively long pulse labeling time was chosen to allow sufficient label incorporation. Heavy cells are harvested immediately after the pulse while medium-heavy and light cells are chased in AHA-free medium for different lengths of time. All three cell populations are then combined and lysed. AHA-containing proteins are purified from the mixed lysate. After digestion into peptides, SILAC-based quantification reveals how much of the pulse-labeled fraction remains at different time points.

We first confirmed that click chemistry-based capturing of heavy AHA-labeled proteins is highly specific and reproducible (Figures S1A–S1C). Next, we checked if AHA might itself affect protein degradation kinetics and found no evidence for this (Figures S1D–S1G). Peptides derived from both the N-terminal and C-terminal halves of AHA-labeled proteins had similar intensities, suggesting that AHA does not induce premature termination (Figure S1H). Collectively, these data indicate that AHA pulse-chase (AHA p-c) enables specific enrichment of newly synthesized proteins with no apparent impact on protein stability, consistent with previous reports (Cohen et al., 2013; Dieterich et al., 2006; Howden et al., 2013; tom Dieck et al., 2015).

For global quantification of protein degradation kinetics, we performed three parallel triple-SILAC experiments with different chase times (1, 2, 4, 8, 16, and 32 hr) in mouse fibroblasts (NIH 3T3). Heavy cells were always harvested immediately after the pulse and served as a common reference point. Exemplary mass spectra for a stable protein (filamin A [Flna]) and an unstable protein (cathepsin L [Ctsl1]) show expected slow and fast degradation, respectively (Figure 1C). The levels of basigin (Bsg) quickly decreased after the chase but then stabilized after ~4 hr. This is consistent with the observation that most of the newly synthesized immature basigin is degraded while the mature form of the protein is stable (Tyler et al., 2012). We combined the data from the three triple-SILAC experiments to obtain kinetic profiles with seven time points. To compensate for differences in cell numbers, we normalized the data using a selected set of very stable proteins (Figure S2; STAR Methods). We also subtracted background signals which could otherwise give rise to erroneous degradation profiles. The entire large-scale experiment was carried out three times, thus yielding data from three independent biological replicates. In total, we obtained profiles for 5,247 proteins (Figure 1D). After applying several quality filters (Figure S3A) we kept 3,605 profiles for further analysis (Figure S3E). Approximately half of these profiles were derived from at least two biological replicates, allowing us to assess reproducibility. Overall, reproducibility was very good with coefficients of variation (CVs, computed in log space) of <10% at time point 32 hr for ~90% of the proteins (Figure S3F).

Stochastic Modeling Reveals Extensive Non-exponential Degradation

To model our experimental protein degradation profiles, we adapted a Markov chain-based approach previously used to study mRNA decay (Figure 2A) (Deneke et al., 2013). We considered two different models. In the first model proteins only exist in a single state "A" that is characterized by a constant decay probability. This "1-state model" therefore describes exponential degradation (ED). The second model has an additional state: newly synthesized proteins first populate state A from where they can either be degraded or transit to state B, which is characterized by a different decay probability. This "2-state model" thus describes non-exponential degradation (NED). To distinguish between both scenarios, we compared the relative quality of both models for each protein degradation profile using the Akaike information criterion (AIC) (Akaike, 1974). This approach considers the trade-off between the goodness of fit and model complexity (that is, the number of parameters). Hence, the 2-state-model is only preferred when the improved fit outweighs the increased complexity. The AIC thus provides a conservative estimate of the fraction of NED proteins. Degradation profiles of Ctsl1 and Flna were better explained by the 1-state model (Figure 2B). In contrast, the degradation profile of basigin was better explained by the 2-state model.

Overall, we found that the profiles of 509 proteins are better explained by the 2-state model (Figure 2C, AIC probability >0.8). This corresponds to ~14% of the 3,605 proteins that passed our quality criteria (Figure S3). We conclude that a sizable number of proteins show a tendency toward NED.

In principle, the 2-state model can describe two different scenarios: when the degradation rate in state A is higher than in state B ($k_A > k_B$) the model describes proteins that become more stable as they age. Alternatively, when the degradation rate in state B is higher than in state A ($k_B > k_A$) the model describes destabilization of older proteins. Intriguingly, we only observed age-dependent stabilization (Figure 2D). This is surprising, as we originally expected proteins to become more unstable over time due to the cumulative effects of age-related damage. However, in our cell line model and within the time period monitored (32 hr) we did not find any evidence for this.

While our AIC probability describes the relative quality of both models, it does not provide information about the extent of NED for individual proteins. We therefore also measured the distance of intermediate data points from the linear fit in the semi log plot (Figure 2E). This delta score (Δ-score) is a simple measure for the extent of non-exponentiality of individual degradation profiles. We then defined NED and ED proteins based on their AIC probabilities and Δ-scores (STAR Methods). With these filters, ~10% of proteins were classified as NED and ~50% as ED. The remaining proteins were classified as undefined ("UN," Figure 2F). For all subsequent analyses, we relied on this classification (Table S1). In summary, our global quantification of cellular protein degradation kinetics reveals that many proteins become more stable once they have survived the first few hours of their existence.

NED Can Be Validated by Independent Methods

Although AHA labeling does not appear to globally affect protein degradation (Figure S1), it is still possible that labeling with an artificial amino acid introduces systematic biases. We therefore wanted to validate our data with independent methods. First, we compared protein degradation rates in the present study with previously published dynamic SILAC data from the same cell line (Schwanhäusser et al., 2011) and found overall good agreement (Figures S4A–S4D). Second, we used classical radioactive pulse-chase and immunoprecipitation to confirm degradation kinetics of an ED and an NED protein (Figure S4E). Finally, to

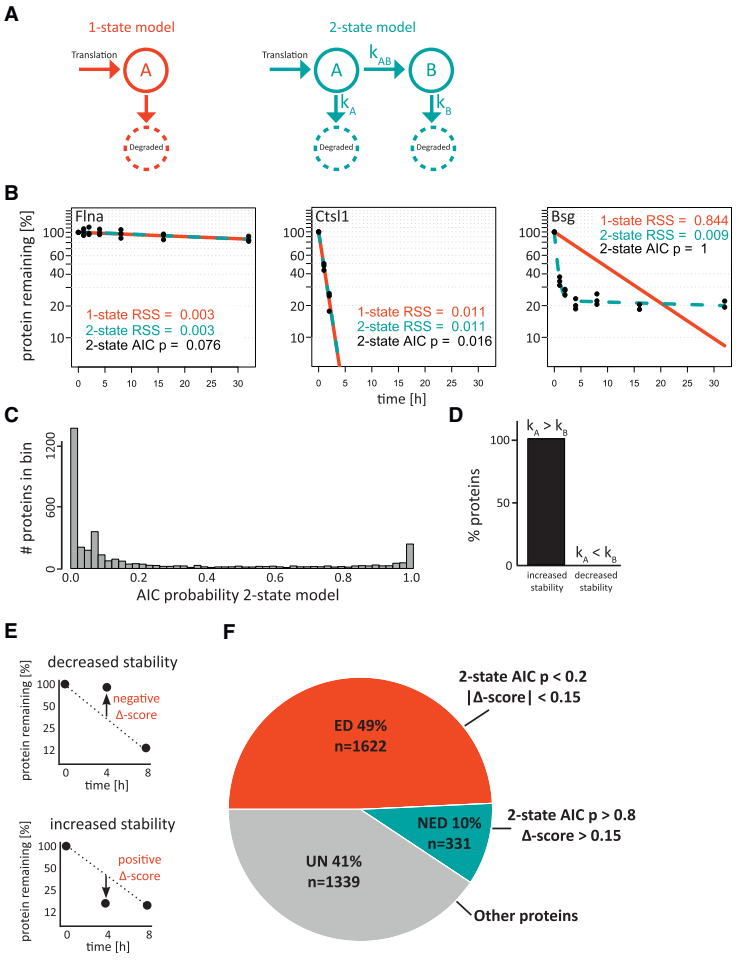


Figure 2. Many Proteins Are Degraded Non-exponentially

(A) Graphical representation of the two Markov models applied. The 1-state model and 2-state model reflect exponential degradation and non-exponential degradation, respectively.

(B) Fitting both models to the exemplary profiles from Figure 1D. For Flna and Ctsl1, both models have residual sum of squares (RSS) of similar size. The Akaike information criterion (AIC) therefore recommends the simpler 1-state model. The profile of Bsg is better explained by the 2-state model.

(C) Histogram of all probabilities for the 2-state model for all proteins that passed our quality criteria.

(D) All proteins with a 2-state probability >0.8 had a larger initial (k_A) than subsequent degradation rate (k_B).

(E) The delta score (Δ-score) as a measure for the extent of non-exponential degradation. For each profile, a straight line is drawn between the 0 and 8 hr time point (in semi log plot). The Δ-score corresponds to the distance of the measurement at 4 hr from this line. Positive and negative Δ-scores indicate age-dependent stabilization and destabilization, respectively.

(F) Fractions of exponentially degraded (ED), non-exponentially degraded (NED), and undefined (UN) proteins defined by their AIC probabilities and Δ-scores.

See also Figure S3.

systematically validate our classification, we designed a novel SILAC-based strategy (Figure 3A): we pulse-labeled light cells with heavy SILAC medium for 4 hr. At this time point, the proteome consists of two populations of protein molecules: light proteins, which are older than 4 hr, and heavy proteins with an age between 0 and 4 hr. We harvested half of the cells to quantify the H/L ratio at this time point. Because NED proteins become more stable with age, the younger heavy protein molecules should degrade faster than the older light ones. We therefore cultivated the other half of cells on medium-heavy growth medium for 8 additional hours. This resulted in the expected decrease in H/L ratios for NED but not for ED proteins (Figures 3B–3D). Label swap experiments and different chase times confirmed this finding (Figures S4F–S4H). We computed a NED validation score based on the average rank of proteins across these SILAC pulse-chase experiments (Figures 3E and 3F). This score indicates how well non-exponential degradation could be validated for individual NED proteins and is included in Table S1. Because these validation experiments have longer pulse times and thus lower sensitivity, a low validation score of a NED protein should not be interpreted as evidence for erroneous classification.

The Ubiquitin Proteasome System Mediates NED

Having shown that a sizable number of proteins are non-exponentially degraded, we next asked how degradation occurs. The two major cellular protein degradation systems are the proteasome and the lysosome. We therefore used drugs, MG132 or wortmannin in combination with bafilomycin A1, to inhibit proteasomal or lysosomal degradation, respectively. To assess non-exponential degradation, we performed AHA p-c experiments with 4 and 8 hr chase times in the presence of the inhibitor or carrier controls (DMSO). The impact of both pathway inhibitions on NED was quantified by measuring their impact on the Δ-score (Figure 4A). Proteasome inhibition reduced Δ-scores of most NED proteins (Figures 4B and 4D; Table S2). In contrast, inhibition of the lysosome did not have an observable impact (Figures 4C and 4E; Table S2). We conclude that the ubiquitin proteasome system is involved in the initial degradation of most NED proteins.

Many NED Proteins Are Members of Multiprotein Complexes

We next characterized features that distinguish ED and NED proteins and found that NED proteins are on average more abundant

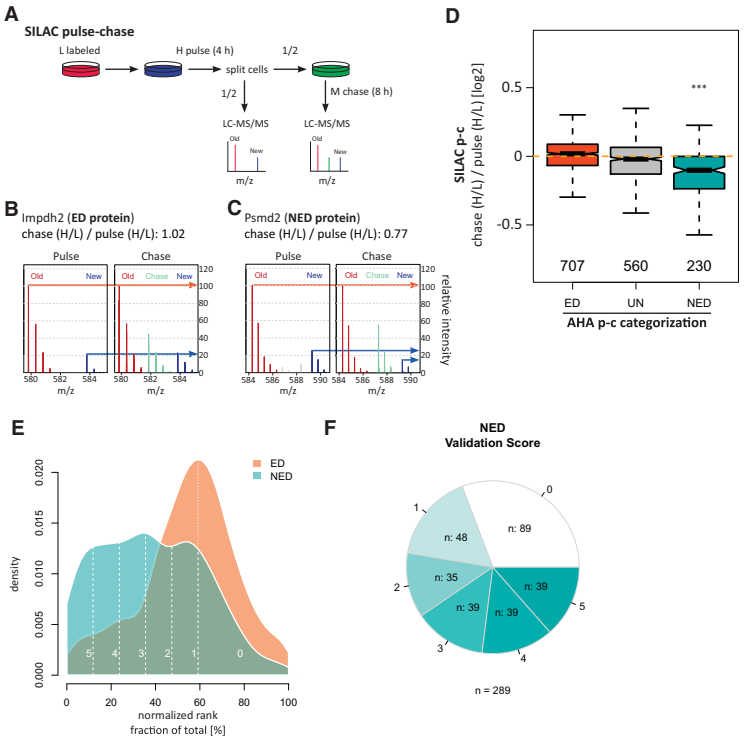


Figure 3. Validation of AHA p-c Data

(A) Experimental design for direct validation of non-exponential decay. Light (L) cells are pulsed with heavy (H) SILAC medium for 4 hr and split into two populations. The first population is harvested immediately while the second is chased for 8 hr in medium-heavy SILAC medium (M chase). If new proteins are less stable than old proteins their H/L ratio is expected to decrease during the chase.

(B and C) Two example spectra from an ED (B) and a NED protein (C) confirm this expectation.

(D) Proteins in the SILAC p-c experiment were classified according to their degradation profile. Only NED proteins show significantly (alpha = 0.05) reduced H/L ratios after the chase compared to all proteins. ***p value <0.0001 from a one-sided Wilcoxon rank-sum test.

(E) Density distributions of ranked SILAC ratios for NED and ED proteins averaged across all four experiments (D and Figures S4F–S4H). NED proteins were assigned a validation score (0 = low validation score; 5 = high validation score) that scales relative to the median of the ED protein distribution.

(F) Counts of NED proteins for the different validation score bins.

See also Figure S4.

and have a higher degree of secondary structure (Figure S5A). Moreover, many NED proteins are members of annotated heteromeric complexes (Figures S5A and S5B). Mapping the data to protein structures showed that 70% of NED proteins are members of heteromeric protein complexes, which is a significant enrichment relative to ED and UN proteins (Figures 5A and S5C; Table S3).

Many complexes contain both NED and ED proteins. Therefore, we investigated if the properties of NED and ED proteins in complexes differ. We found that NED proteins have significantly larger interaction interfaces within complexes (Figure 5B; Table S3). It has been shown that complexes form via ordered and evolutionarily conserved assembly pathways (Marsh et al., 2013). We found that NED subunits assemble significantly earlier than ED subunits (Figure 5C; Table S3). Finally, we found that the NED subunits show stronger correlated coexpression with other subunits of the same complex, while the ED subunits show less coherent expression (Figures 5D and S5D; Table S3). Together, these data indicate that NED proteins are typically core components of multiprotein complexes while ED proteins tend to be more peripheral.

A long-standing hypothesis is that proteins are stabilized by complex formation (Goldberg, 2003). While several individual examples support this idea (Blikstad et al., 1983; Lam et al., 2007; Toyama et al., 2013) it has, to the best of our knowledge, not yet been investigated systematically. We reasoned that complex formation could explain non-exponential degradation (Figure 5E): if NED proteins were produced in excess relative to ED proteins, the degradation of the non-assembled subunits would lead to non-exponential decay. We therefore analyzed the abundance of newly synthesized proteins directly after the pulse. To this end, we normalized the data in a complex-centric manner (Figure 5F). We found that NED proteins are indeed over-synthesized relative to ED proteins in the same complex. Moreover, while the initial degradation rates (i.e., the degradation rates of state A) of NED proteins within a complex varied considerably, their second (state B) degradation rates were more similar and close to the degradation rates of the ED subunits (Figure S6). Thus, many NED proteins in complexes appear to be produced in super-stoichiometric amounts relative to their ED counterparts.

To independently validate these findings, we performed the same complex-centric analysis on ribosome profiling data (Subtelny et al., 2014) and observed the same trend (Figure 5G). To directly assess the link between complex assembly and NED we focused on the ribosome—a complex rich in NED proteins. Because the ribosome consists of proteins and rRNAs, ribosome assembly can be inhibited by blocking rRNA transcription. We found that actinomycin D treatment increased the initial degradation of ribosomal proteins (Figure 5H), consistent with our hypothesis and with previous data (Lam et al., 2007).

NED Is Evolutionarily Conserved

All data presented so far are based on the analysis of mouse fibroblasts (NIH 3T3). We therefore asked if our findings are due to specific features of this model system. For example, even though NIH 3T3 cells are derived from primary cells, a recent cytogenetic study revealed a complex rearranged karyotype

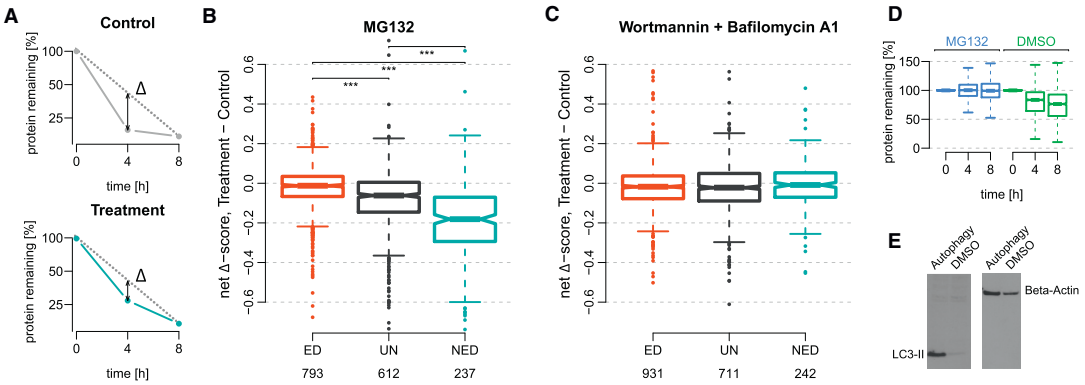


Figure 4. NED Is Decreased by Proteasome Inhibition

(A) To quantify the impact of inhibitors on non-exponential degradation the Δ-score in treated and control cells is compared.

(B and C) Distributions of net Δ-scores for ED, UN, and NED proteins displayed as boxplots. The proteasome inhibitor MG132 significantly reduced Δ-scores of NED proteins (one-sided Wilcoxon rank-sum test; ***$p < 0.0001$) while the autophagy inhibitors wortmannin and bafilomycin A had no significant impact (alpha = 0.05). Numbers of proteins in each group are depicted. Classification of proteins is based on the original AHA p-c experiment (Figure 2F).

(D) We estimated the effect of MG132 on general protein degradation by plotting "% protein remaining" for all proteins with or without treatment. MG132 stabilized the majority of the measured proteins.

(E) To control for inhibition of lysosomal degradation, samples acquired in parallel to the MS experiment were analyzed by western blot and probed against the autophagy marker LC3-II and for β-actin. "Autophagy" in the plot refers to the inhibitor combo.

(Leibiger et al., 2013). It is therefore possible that the super-stoichiometric synthesis of NED proteins is due to genomic amplification of the corresponding genes. In this case, our data would have little relevance for other model systems.

To test this possibility, we analyzed the diploid human retinal pigmented epithelial cell line RPE-1. Low coverage whole genome sequencing of this cell line confirmed that, with the exception of partial trisomies for chromosome 10 and 12, it is mostly diploid (Figure S7). We performed three independent large AHA p-c experiments to obtain degradation profiles for 4,079 proteins. A total of 47% and 9.4% of the proteins that passed the quality filters (n = 3,133) were classified as ED and NED, respectively (Figure 6A, left bar; Table S4). These fractions are similar to the mouse fibroblast data. We then grouped the human proteins according to the degradation profiles of their mouse orthologs (Figure 6A). Human orthologs of mouse NED proteins were enriched in NED proteins. Conversely, human orthologs of mouse ED proteins were enriched in ED proteins. In addition, the Δ-scores of the human proteins and their mouse orthologs were significantly correlated (Figure 6B). This is surprising, especially because there are many reasons why degradation profiles may actually differ (cell-cycle properties, overall proteome composition that may drastically differ between both cell types). We also found that human NED proteins that are part of multiprotein complexes tend to be produced in super-stoichiometric amounts (Figure 6C), consistent with the mouse data (Figure 5F). This observation still holds when human proteins are classified according to the degradation profile of their mouse orthologs, further supporting the notion of conservation.

To further assess the conservation of NED, we used ribosome profiling data from different species (human, HEK293 [Liu et al., 2013], mouse [Shalgi et al., 2013], zebrafish [Chew et al., 2013], and *Caenorhabditis elegans* [Nedialkova and Leidel, 2015]). We classified proteins in these datasets according to the degradation profiles of their human orthologs (in RPE-1 cells). We then normalized the ribosome protected fragment (RPF) reads in a complex-centric manner. We found that orthologs of human NED proteins tend to be over-synthesized in mouse and zebrafish (Figure 6D). The trend also holds in *C. elegans* although it is not significant. Together, these data show that non-exponential degradation is not mainly due to "erroneous" protein overproduction caused by genomic rearrangements in a specific cell line. Instead, our findings show that protein degradation kinetics are—at least partially—conserved between species and independent of the cell type.

NED Predicts Protein Level Attenuation in Aneuploidy

Based on our findings, we propose a simple model that explains the relationship between protein degradation kinetics and complex formation (Figure 7A). Accordingly, NED proteins are overproduced relative to the ED proteins in the same complex. Therefore, only a fraction of the overproduced proteins are stabilized by complex formation while the rest are degraded. Importantly, this overproduction occurs in disomic cells and is thus not generally due to aneuploidy. However, we reasoned that aneuploid cells would allow us to test the model: genomic amplification of NED proteins should increase their over-production and thus the initial degradation (Figure 7A). Consequently, amplification of genes encoding NED proteins should not lead to correspondingly increased protein levels. Instead, NED proteins should be attenuated, i.e., their protein levels should remain relatively constant despite the genomic amplification (Dephoure et al., 2014; Geiger et al., 2010; Stingele et al., 2012).

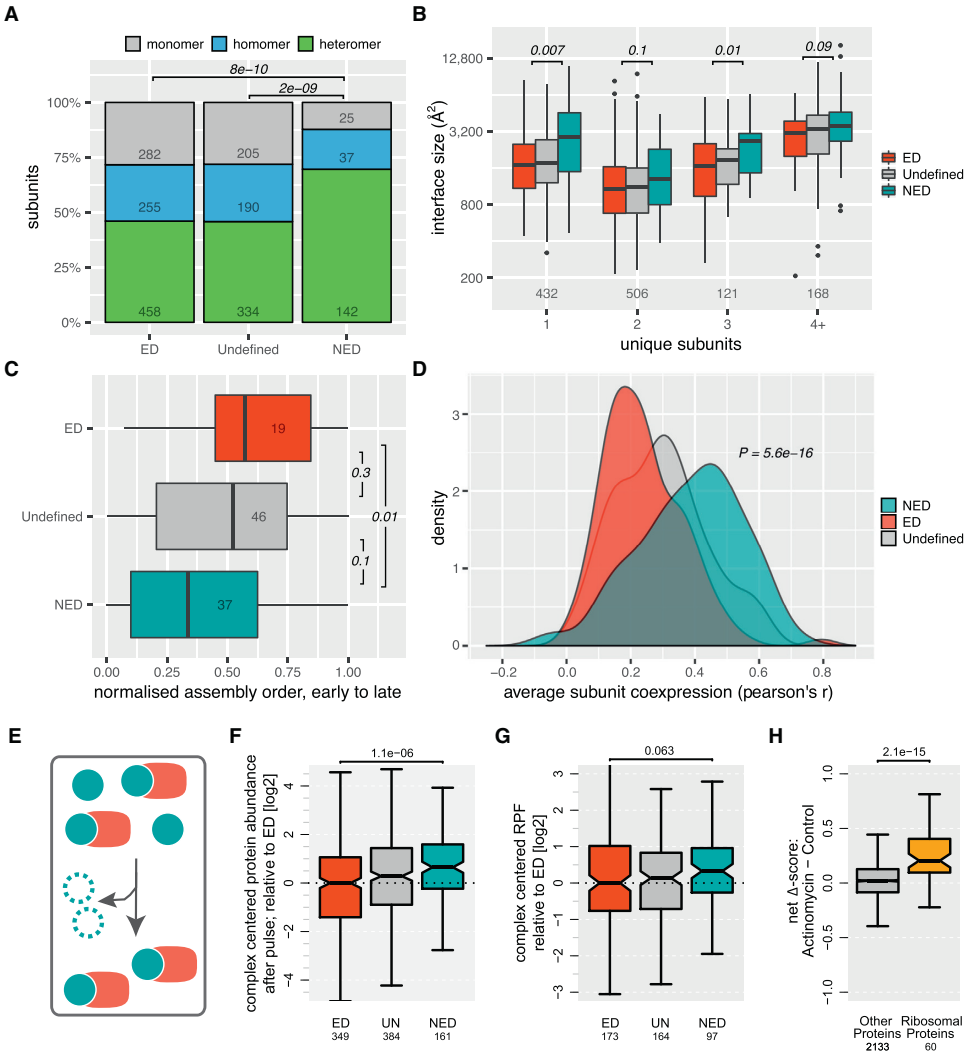


Figure 5. NED and Protein Complexes

(A) NED proteins are significantly overrepresented in heteromeric protein complexes (Fisher's exact test, heteromeric versus monomeric or homomeric subunits). Numbers within bars represent raw subunit counts. The trend holds when ribosomal proteins are excluded (Figure S6C).

(B) NED proteins tend to form larger interfaces in complexes than ED proteins. Subunits were binned by the number of unique subunits per complex to control for the fact that NED proteins are overrepresented in larger complexes. P values were calculated using Wilcoxon rank-sum tests, comparing NED to ED. Subunit counts are given along the bottom.

(C) NED subunits of large complexes (greater than five unique subunits) tend to assemble earlier than ED subunits. Normalized assembly scores of 0-to-1 indicate the first-to-last steps of a given (dis)assembly pathway. P values were calculated using Wilcoxon rank-sum tests. Raw subunit counts are given within each box. The difference in assembly order was not significant for smaller complexes.

(D) NED proteins show stronger coexpression (mRNA level). For each subunit, the average coexpression correlation coefficient is calculated with all other subunits within the same complex. P value is calculated with the Wilcoxon rank-sum test comparing NED to ED.

(E) A simple model can explain non-exponential degradation of subunits of a heteromeric complex: NED proteins (turquoise) are synthesized in super-stoichiometric amounts relative to ED proteins (red). Only a fraction of the NED protein molecules is stabilized by complex formation while the excess is degraded.

(F) NED proteins tend to be produced in super-stoichiometric amounts. Protein abundances after the pulse (t = 0 hr) were normalized in a complex centered manner. p values are based on one-sided Wilcoxon rank-sum tests.

(legend continued on next page)

To test this prediction, we took advantage of RPE-1 cells that were engineered to carry one additional copy of specific chromosomes (Stingele et al., 2012). Low coverage genome sequencing (Figures S7A and S7B) and chromosome painting (Figures S7C and S7D) verified that these cells are trisomic for chromosome 5 and part of chromosome 11. We therefore also performed AHA p-c experiments with these "RPE-1 trisomic" cells (Table S5). We then compared protein production in RPE-1 and RPE-1 trisomic cells using abundance levels after the pulse as a proxy. As expected, proteins encoded by trisomic regions were upregulated in trisomic cells (Figure 7B). Note that chromosome 10 and 12 were also partially trisomic in the parental cell line and therefore excluded from subsequent analyses.

We next compared the extent of non-exponential degradation in RPE-1 and RPE-1 trisomic cells using the Δ-score. Proteins encoded in disomic regions of the genome had similar Δ-scores in both cell lines (Figure 7C). However, consistent with our prediction, NED proteins in trisomic regions displayed significantly increased non-exponential degradation as measured by the change in their Δ-scores. Importantly, this behavior was specific for NED proteins and not observed for ED proteins. We conclude that increasing over-production of NED proteins tends to increase their initial degradation.

Aneuploidy has severe developmental effects, it is the leading cause of mental retardation and spontaneous abortions, as well as a hallmark of cancer (Santaguida and Amon, 2015). However, the functional consequences of aneuploidy are only beginning to emerge. Studies in yeast and mammalian cell lines have shown that mRNA levels generally scale with gene copy numbers. However, protein levels are sometimes attenuated toward the euploid state. It has also been noted that most attenuated proteins are members of multiprotein complexes (Dephoure et al., 2014; Geiger et al., 2010; Stingele et al., 2012). However, not all proteins in multiprotein complexes are attenuated and not all proteins that are attenuated are part of (known) multiprotein complexes.

Our model predicts that NED proteins should be attenuated. To test this idea, we directly quantified relative changes in steady-state protein levels in RPE-1 and RPE-1 trisomic cells in a separate SILAC experiment. We found that NED proteins encoded in trisomic regions were more attenuated than ED proteins (Figure 7D). This is consistent with our model, even though interpreting the data is complicated by the rather small number of NED proteins in trisomic regions. We also observed this effect in the subset of proteins that are part of annotated complexes (Figure 7E). Hence, protein degradation kinetics can explain why some proteins in complexes show more attenuation (NED proteins) than others (ED proteins). Moreover, even for the subset of proteins that are not part of an annotated complex, NED proteins showed more attenuation than ED proteins (Figure 7F). Collectively, these data show that NED can help to predict protein level attenuation in aneuploidy.

DISCUSSION

Our kinetic analysis of protein stability reveals widespread age-dependent degradation. We find that kinetic profiles are similar in two different cell types and conserved between mouse and humans. Many non-exponentially degraded proteins are subunits of multiprotein complexes that are produced in excess. Accordingly, we propose a simple model in which only a fraction of newly synthesized NED proteins is stabilized by complex formation while the rest are degraded. This model can help to predict how changes in DNA copy-number affect protein levels.

While the AHA p-c method employed here has many advantages, it is also important to keep its limitations in mind. For example, due to the technical challenges of AHA p-c, our data are not comprehensive. We do not know how our findings extrapolate to the uncovered part of the proteome. Also, the pulse time of 1 hr is too long to capture events on the timescale of minutes. Therefore, our data cannot be used to estimate the extent of co- or peri-translational degradation (Duttler et al., 2013; Schubert et al., 2000; Wheatley et al., 1980). Moreover, because our longest chase time is 32 hr, we cannot tell what happens to older proteins. It is possible that longer chase times would reveal age-dependent destabilization. Additionally, the seven time points we covered allowed us to distinguish between a 1-state and a 2-state model, whereas more complex multistage models would require measurements at more time points. Increasing the number of data points per profile would also decrease the number of proteins we could not classify as NED or ED. We classified these unclear cases as "undefined"—a purely technical definition that has probably no biological significance. Finally, despite validation by independent methods, we cannot fully rule out biases in our AHA p-c data.

We find that some proteins become very stable after they have survived the first few hours of their molecular life. We have experienced that this surprises many scientists because they intuitively believe that all proteins are constantly turned over. However, our finding is consistent with other observations. For example, SILAC labeling of post-mitotic cells typically remains incomplete despite long labeling times (Liao et al., 2008). Moreover, using metabolic labeling of rats, Toyama et al. (2013) identified several proteins with extraordinarily long lifespans in the rat brain. The latter study also provided evidence that the high stability of one protein, Nup96, is due to its deposition into a stable complex.

The observation that mammalian cells overproduce specific subunits of multiprotein complexes is surprising and in marked contrast to *Escherichia coli* where subunits were reported to be made in precise proportion to their required stoichiometry (Li et al., 2014). Why are mammalian cells producing excess amounts of specific proteins? We would like to discuss four potential answers to this question:

(G) Complex-centered analysis of ribosome profiling data supports super-stoichiometric production of NED proteins. p values are based on one-sided Wilcoxon rank-sum tests.
(H) Inhibition of rRNA synthesis with actinomycin D selectively increased Δ-scores of ribosomal proteins. See Figure 4A for experimental design. P values are based on one-sided Wilcoxon rank-sum tests.
See also Figure S5 and Table S3.

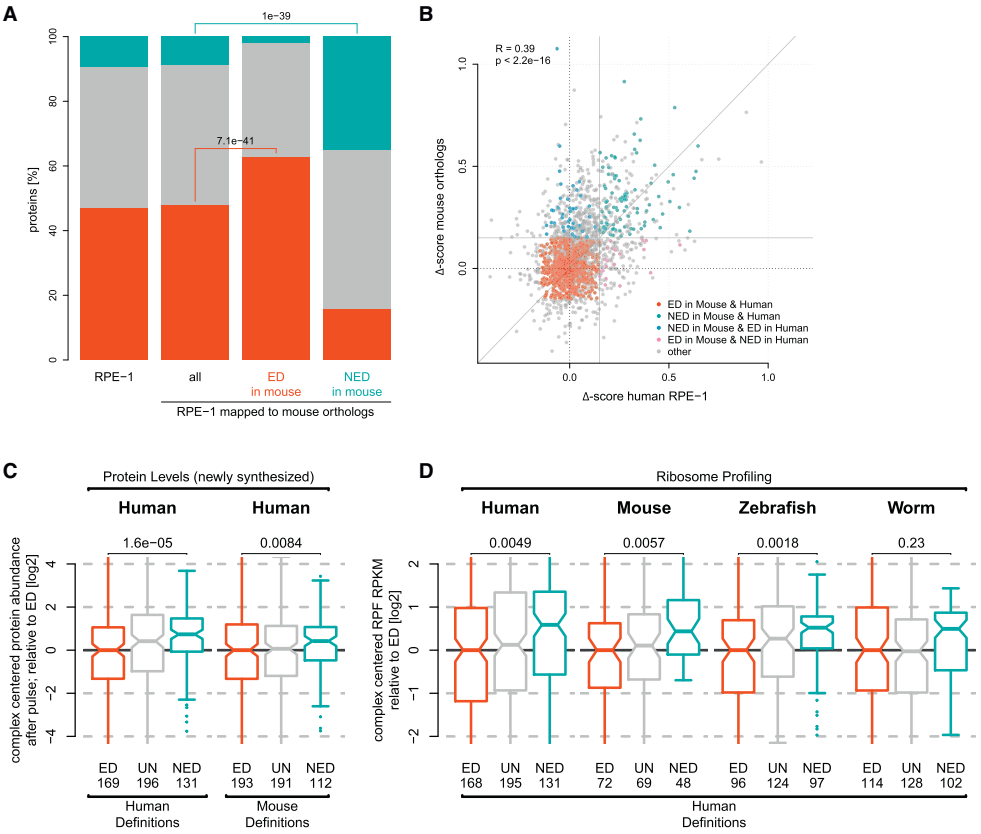


Figure 6. NED Is Evolutionarily Conserved

(A) Relative fractions of NED (turquoise), ED (red), and undefined (gray) proteins in the diploid human epithelial cell line RPE-1. We mapped human proteins to their mouse orthologs and grouped them according to their degradation profile in mouse fibroblasts. Human proteins with ED mouse orthologs are enriched in ED proteins. Similarly, human proteins with NED mouse orthologs are enriched in NED proteins. P values are based on a hypergeometric test.

(B) Orthologous human and mouse proteins show significantly correlated Δ-scores. Pearson's correlation coefficient (R) is derived from all plotted Δ-score pairs.

(C) NED subunits of protein complexes are synthesized in super-stoichiometric amounts relative to other subunits in human (left). This is even the case when the mouse definitions (for ED, UN, and NED) are used on the human dataset (right).

(D) Analysis of ribosome profiling data from several species confirms that the super-stoichiometric synthesis of NED proteins is evolutionarily conserved. Depicted p values are based on one-sided Wilcoxon rank-sum tests.

See also Figure S7.

(1) It is possible that only a fraction of newly synthesized proteins adopt a functional state while the rest are terminally misfolded and thus degraded. However, because NED proteins tend to be relatively short and well-structured (Figure S5A) this explanation does not seem to generally hold.

(2) The observation that ED proteins tend to be disordered (Figure S5A) suggests an alternative explanation: disordered proteins are often harmful when overexpressed, which is probably due to their tendency to make promiscuous interactions (Vavouri et al., 2009). Overproduction of NED proteins may have evolved to ensure that potentially harmful ED proteins are never alone: The super-stoichiometric synthesis of "benign" NED subunits would be a failsafe mechanism against deleterious effects of unbalanced production of the more harmful ED proteins. NED proteins would thus have a chaperone-like function toward their cognate ED proteins.

(3) ED proteins might have evolved as limiting factors to facilitate the coordinated regulation of protein complex abundance: Upregulating an ED protein would stabilize interacting NED proteins and thus increase the abundance of the entire complex. Conversely, reducing the expression of the limiting ED protein would decrease complex abundance. This explanation resonates well with the finding that mRNAs encoding ED proteins show less co-expression with mRNAs encoding other subunits and have longer 5′ and 3′ UTRs (Figure 5D and data not

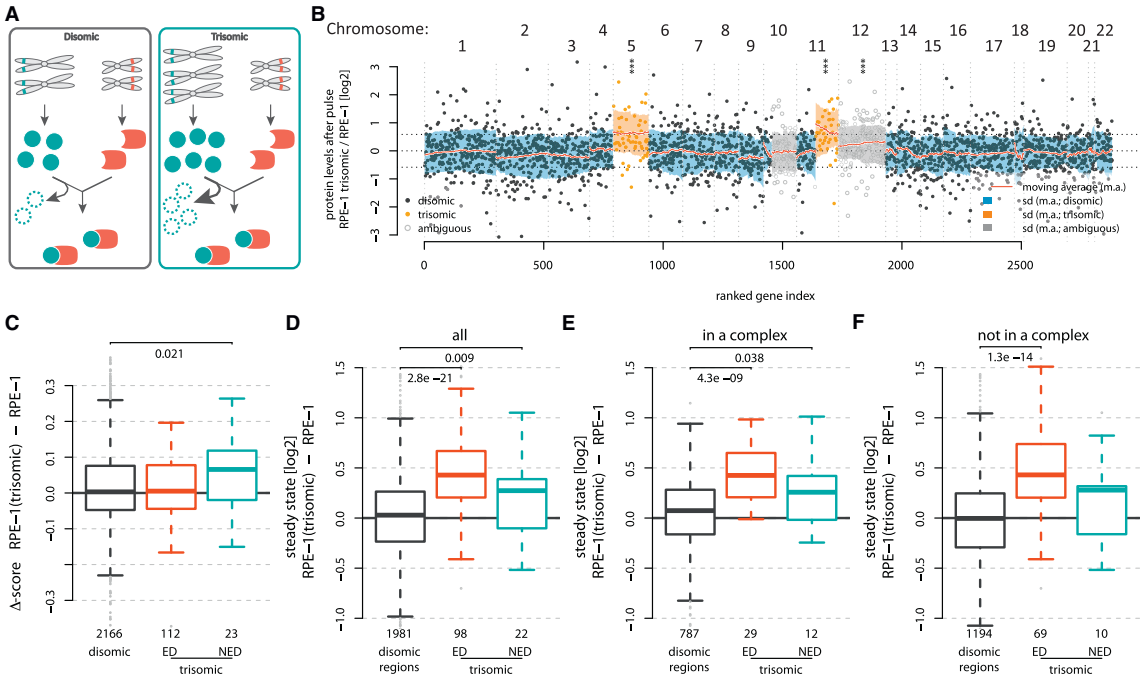


Figure 7. NED, Aneuploidy, and Attenuation

(A) A model depicting the expected impact of gene amplification on protein synthesis and degradation. In normal (that is, disomic) cells NED proteins are over-synthesized relative to ED proteins in the same complex. Degradation of the excess molecules gives rise to their NED profile. Genomic amplification of NED proteins further increases over-production and thus initial degradation.

(B) Log2 fold changes of protein abundances after pulse in RPE-1 cells and RPE-1 cells carrying extra copies of specific chromosomes (sorted by chromosome and genomic position). The data were divided into disomic (black), trisomic (orange) and ambiguous regions (gray) based on genome sequencing data (Figure S7). Regions with significantly different protein abundance in comparison to the disomic cells are marked with asterisks (one-sided Wilcoxon rank-sum; ***$p < 0.0001$).

(C) NED proteins show increased initial degradation (Δ-scores) when the corresponding genes are in trisomic regions. Degradation of ED proteins is not affected.

(D–F) NED predicts protein level attenuation. We compared steady-state protein levels in RPE-1 and RPE-1 trisomic cells using standard SILAC. Boxplots show log2 fold changes for ED and NED proteins in trisomic regions compared to all proteins in disomic regions. This analysis is shown for all proteins (D), proteins that are part of complexes (E), and proteins that are not part of complexes (F). The number of analyzable protein pairs is displayed below each boxplot. P values were computed using one-sided Wilcoxon rank-sum tests and are shown for significantly different distributions (alpha = 0.05).

See also Figure S7.

shown). It also fits to data from yeast showing that most complexes consist of both constitutive and periodically expressed subunits (de Lichtenberg et al., 2005) and to the finding that complex stoichiometry can vary across tissues in mammals (Ori et al., 2016).

(4) Overproduction of NED proteins may be important for ordered complex assembly (Marsh et al., 2013; Matalon et al., 2014): because the formation of protein-protein interactions depends on protein concentration, the relative abundance of proteins may in part determine the assembly order. This explanation is consistent with our observation that NED proteins tend to assemble earlier (Figure 5C).

Which (if any) of these four possibilities explains protein over-production and NED remains to be investigated. It is likely that different reasons are relevant for different proteins. It is also important to note that not all NED proteins are components of (known) multiprotein complexes. NED of monomers could be due to many different molecular mechanisms. For example, NED of CFTR and basigin appears to be due to failed protein folding in the ER (Tyler et al., 2012; Ward and Kopito, 1994). More generally, biphasic degradation can be due to the existence of distinct pools of a protein. For example, these pools may reflect residency in different compartments (cytosol, nucleus, mitochondria, extracellular, etc.), different posttranslational modification states (glycosylation, phosphorylation, etc.), or assembly into different complexes. While our manuscript focuses on the latter possibility, this is by no means the only possible explanation. Finally, while we have interpreted the age-dependent rate as giving information on the aging of every individual molecule, we would like to note that an alternative interpretation based on frailty theory (Aalen, 1994) would give similar fitting quality.

Our results have significant implications for aneuploidy. First, we confirm the previous observation that amplified genes encoding members of multiprotein complexes are often attenuated at the protein level (Dephoure et al., 2014; Geiger et al., 2010; Stingele et al., 2012). This finding was interpreted in the light of the longstanding idea that unassembled subunits of complexes are unstable (Goldberg, 2003). Accordingly, overproduction caused by gene amplification is attenuated for proteins in complexes. Our findings considerably extend this concept. The important difference is that we already observe unbalanced subunit production at baseline (Figure 7A). Consequently, our model can explain why only some subunits of complexes show attenuation (Figure 7E). More broadly, our data help to predict how protein levels change in response to altered protein production. We expect that this will turn out to be useful for comprehending the complex cellular phenotypes of aneuploidy and somatic copy-number alterations in cancer. The data might also help to explain posttranslational buffering (Battle et al., 2015).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2016.09.015.

AUTHOR CONTRIBUTIONS

E.M. performed most wet lab experiments and contributed to study design, data analysis, and data interpretation. C.S. and A.V. developed and implemented the normalization, mathematical model, model calibration, and model selection procedure. H.Z. performed most of the bioinformatic analyses such as validation score, complex-centered, AUC, aneuploidy, ribosome profiling, and conservation analysis. J.N.W. and J.A.M. analyzed the complex assembly data (Figures 5A–5D). N.D. and Z.S. generated the trisomic RPE-1 cell line, performed chromosome paintings, and helped in interpreting the aneuploidy data. X.W., J.H., and W.C. sequenced RPE-1 and RPE-1 (trisomic) cells and computed copy-number estimates. M.S. conceived and supervised the study and wrote the manuscript with input from all authors.

ACKNOWLEDGMENTS

We thank Katrin Eichelbaum for help with establishing AHA labeling, Piotr Grabowski for setting up SCX chromatography, and Koshi Imami (all Max Delbrück Center for Molecular Medicine [MDC]) for establishing RP-HPLC on monolithic columns. We also thank Thomas Sommer and his lab members (MDC) as well as Thomas Langer and Simon Tröder (University of Cologne) for fruitful discussions. Christian Sommer (MDC), Eva Kärgel (MDC), and Yvonne Kraus (Max Planck Institute of Biochemistry) provided excellent technical assistance and Rebecca Eccles (MDC) helped proofreading the manuscript. This work was in part funded by a grant of the Helmholtz Association to M.S. and by a grant of the European Union (ITN "NICHE") to C.S. and A.V. J.M. is supported by a Medical Research Council Career Development Award (MR/M02122X/1).

Received: March 30, 2016
Revised: July 19, 2016
Accepted: September 7, 2016
Published: October 6, 2016

REFERENCES

Aalen, O.O. (1994). Effects of frailty in survival analysis. Stat. Methods Med. Res. *3*, 227–243.

Aalen, O.O., and Gjessing, H.K. (2001). Understanding the shape of the hazard rate: a process point of view (With comments and a rejoinder by the authors). Stat. Sci. *16*, 1–22.

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions Autom. Control *19*, 716–723.

Andersen, J.S., Lam, Y.W., Leung, A.K., Ong, S.E., Lyon, C.E., Lamond, A.I., and Mann, M. (2005). Nucleolar proteome dynamics. Nature *433*, 77–83.

Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., and Gilad, Y. (2015). Genomic variation. Impact of regulatory variation from RNA to protein. Science *347*, 664–667.

Blikstad, I., Nelson, W.J., Moon, R.T., and Lazarides, E. (1983). Synthesis and assembly of spectrin during avian erythropoiesis: stoichiometric assembly but unequal synthesis of alpha and beta spectrin. Cell *32*, 1081–1091.

Bourdetsky, D., Schmelzer, C.E., and Admon, A. (2014). The nature and extent of contributions by defective ribosome products to the HLA peptidome. Proc Natl Acad Sci USA *111*, 1591–1599.

Burnham, K.P., and Anderson, D.R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (Springer-Verlag).

Chew, G.L., Pauli, A., Rinn, J.L., Regev, A., Schier, A.F., and Valen, E. (2013). Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. Development *140*, 2828–2834.

Ciechanover, A. (2005). Aaron Ciechanover - Nobel Lecture: Intracellular Protein Degradation, The Nobel Prizes 2004 (Science History Publications/ Watson Publishing).

Cohen, L.D., Zuchman, R., Sorokina, O., Müller, A., Dieterich, D.C., Armstrong, J.D., Ziv, T., and Ziv, N.E. (2013). Metabolic turnover of synaptic proteins: kinetics, interdependencies and implications for synaptic maintenance. PLoS ONE *8*, e63191.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.

de Lichtenberg, U., Jensen, L.J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. Science *307*, 724–727.

Deneke, C., Lipowsky, R., and Valleriani, A. (2013). Complex degradation processes lead to non-exponential decay patterns and age-dependent decay rates of messenger RNA. PLoS ONE *8*, e55442.

Dephoure, N., Hwang, S., O'Sullivan, C., Dodgson, S.E., Gygi, S.P., Amon, A., and Torres, E.M. (2014). Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. eLife *3*, e03023.

Dieterich, D.C., Link, A.J., Graumann, J., Tirrell, D.A., and Schuman, E.M. (2006). Selective identification of newly synthesized proteins in mammalian cells using bioorthogonal noncanonical amino acid tagging (BONCAT). Proc. Natl. Acad. Sci. USA *103*, 9482–9487.

Doherty, M.K., Hammond, D.E., Clague, M.J., Gaskell, S.J., and Beynon, R.J. (2009). Turnover of the human proteome: determination of protein intracellular stability by dynamic SILAC. J. Proteome Res. *8*, 104–112.

Duttler, S., Pechmann, S., and Frydman, J. (2013). Principles of cotranslational ubiquitination and quality control at the ribosome. Mol. Cell *50*, 379–393.

Eichelbaum, K., and Krijgsveld, J. (2014). Rapid temporal dynamics of transcription, protein synthesis, and secretion during macrophage activation. Mol. Cell. Proteomics *13*, 792–810.

Eichelbaum, K., Winter, M., Berriel Diaz, M., Herzig, S., and Krijgsveld, J. (2012). Selective enrichment of newly synthesized proteins for quantitative secretome analysis. Nat. Biotechnol. *30*, 984–990.

Eravci, M., Sommer, C., and Selbach, M. (2014). IPG strip-based peptide fractionation for shotgun proteomics. Methods Mol. Biol. *1156*, 67–77.

Geiger, T., Cox, J., and Mann, M. (2010). Proteomic changes resulting from gene copy number variations in cancer cells. PLoS Genet. *6*, e1001090.

Goldberg, A.L. (2003). Protein degradation and protection against misfolded or damaged proteins. Nature *426*, 895–899.

Goldberg, A.L., and Dice, J.F. (1974). Intracellular protein degradation in mammalian and bacterial cells. Annu. Rev. Biochem. *43*, 835–869.

Hinkson, I.V., and Elias, J.E. (2011). The dynamic state of protein turnover: It's about time. Trends Cell Biol. *21*, 293–303.

Hou, J., Wang, X., McShane, E., Zauber, H., Sun, W., Selbach, M., and Chen, W. (2015). Extensive allele-specific translational regulation in hybrid mice. Mol. Syst. Biol. *11*, 825.

Howden, A.J.M., Geoghegan, V., Katsch, K., Efstathiou, G., Bhushan, B., Boutureira, O., Thomas, B., Trudgian, D.C., Kessler, B.M., Dieterich, D.C., et al. (2013). QuaNCAT: quantitating proteome dynamics in primary cells. Nat. Methods *10*, 343–346.

Jovanovic, M., Rooney, M.S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E.H., Fields, A.P., Schwartz, S., Raychowdhury, R., et al. (2015). Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. Science *347*, 1259038.

Kiick, K.L., Saxon, E., Tirrell, D.A., and Bertozzi, C.R. (2002). Incorporation of azides into recombinant proteins for chemoselective modification by the Staudinger ligation. Proc. Natl. Acad. Sci. USA *99*, 19–24.

Kim, W., Bennett, E.J., Huttlin, E.L., Guo, A., Li, J., Possemato, A., Sowa, M.E., Rad, R., Rush, J., Comb, M.J., et al. (2011). Systematic and quantitative assessment of the ubiquitin-modified proteome. Mol. Cell *44*, 325–340.

Kristensen, A.R., Gsponer, J., and Foster, L.J. (2013). Protein synthesis rate is the predominant regulator of protein expression during differentiation. Mol. Syst. Biol. *9*, 689.

Kulak, N.A., Pichler, G., Paron, I., Nagaraj, N., and Mann, M. (2014). Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. Nat. Methods *11*, 319–324.

Lam, Y.W., Lamond, A.I., Mann, M., and Andersen, J.S. (2007). Analysis of nucleolar protein dynamics reveals the nuclear degradation of ribosomal proteins. Curr. Biol. *17*, 749–760.

Larance, M., Ahmad, Y., Kirkwood, K.J., Ly, T., and Lamond, A.I. (2013). Global subcellular characterization of protein degradation using quantitative proteomics. Mol. Cell. Proteomics *12*, 638–650.

Leibiger, C., Kosyakova, N., Mkrtchyan, H., Glei, M., Trifonov, V., and Liehr, T. (2013). First molecular cytogenetic high resolution characterization of the NIH 3T3 cell line by murine multicolor banding. J. Histochem. Cytochem. *61*, 306–312.

Li, G.W., Burkhardt, D., Gross, C., and Weissman, J.S. (2014). Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell *157*, 624–635.

Liao, L., Park, S.K., Xu, T., Vanderklish, P., and Yates, J.R., 3rd. (2008). Quantitative proteomic analysis of primary neurons reveals diverse changes in synaptic protein content in fmr1 knockout mice. Proc. Natl. Acad. Sci. USA *105*, 15281–15286.

Liu, B., Han, Y., and Qian, S.B. (2013). Cotranslational response to proteotoxic stress by elongation pausing of ribosomes. Mol. Cell *49*, 453–463.

Marsh, J.A., Hernández, H., Hall, Z., Ahnert, S.E., Perica, T., Robinson, C.V., and Teichmann, S.A. (2013). Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell *153*, 461–470.

Matalon, O., Horovitz, A., and Levy, E.D. (2014). Different subunits belonging to the same protein complex often exhibit discordant expression levels and evolutionary properties. Curr. Opin. Struct. Biol. *26*, 113–120.

Motoyama, A., Xu, T., Ruse, C.I., Wohlschlegel, J.A., and Yates, J.R., 3rd. (2007). Anion and cation mixed-bed ion exchange for enhanced multidimensional separations of peptides and phosphopeptides. Anal. Chem. *79*, 3623–3634.

Nedialkova, D.D., and Leidel, S.A. (2015). Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity. Cell *161*, 1606–1618.

Okamura, Y., Aoki, Y., Obayashi, T., Tadaka, S., Ito, S., Narise, T., and Kinoshita, K. (2015). COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. Nucleic Acids Res. *43*, D82–D86.

Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol. Cell. Proteomics *1*, 376–386.

Ori, A., Iskar, M., Buczak, K., Kastritis, P., Parca, L., Andrés-Pons, A., Singer, S., Bork, P., and Beck, M. (2016). Spatiotemporal variation of mammalian protein complex stoichiometries. Genome Biol. *17*, 47.

Puchades, M., Westman, A., Blennow, K., and Davidsson, P. (1999). Removal of sodium dodecyl sulfate from protein samples prior to matrix-assisted laser desorption/ionization mass spectrometry. Rapid Commun. Mass Spectrom. *13*, 344–349.

Rappsilber, J., Ishihama, Y., and Mann, M. (2003). Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal. Chem. *75*, 663–670.

Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O.N., Stümpflen, V., and Mewes, H.W. (2008). CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res. *36*, D646–D650.

Santaguida, S., and Amon, A. (2015). Short- and long-term effects of chromosome mis-segregation and aneuploidy. Nat. Rev. Mol. Cell Biol. *16*, 473–485.

Schimke, R.T., and Doyle, D. (1970). Control of enzyme levels in animal tissues. Annu. Rev. Biochem. *39*, 929–976.

Schoenheimer, R. (1942). The Dynamic State of Body Constituents (Harvard University Press).

Schubert, U., Antón, L.C., Gibbs, J., Norbury, C.C., Yewdell, J.W., and Bennink, J.R. (2000). Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. Nature *404*, 770–774.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58–63.

Shalgi, R., Hurt, J.A., Krykbaeva, I., Taipale, M., Lindquist, S., and Burge, C.B. (2013). Widespread regulation of translation by elongation pausing in heat shock. Mol. Cell *49*, 439–452.

Sheean, M.E., McShane, E., Cheret, C., Walcher, J., Müller, T., Wulf-Goldenberg, A., Hoelper, S., Garratt, A.N., Krüger, M., Rajewsky, K., et al. (2014). Activation of MAPK overrides the termination of myelin growth and replaces Nrg1/ErbB3 signals during Schwann cell development and myelination. Genes Dev. *28*, 290–303.

Sin, C., Chiarugi, D., and Valleriani, A. (2016). Degradation parameters from pulse-chase experiments. PLoS ONE *11*, e0155028.

Sormanni, P., Camilloni, C., Fariselli, P., and Vendruscolo, M. (2015). The s2D method: simultaneous sequence-based prediction of the statistical populations of ordered and disordered regions in proteins. J. Mol. Biol. *427*, 982–996.

Stingele, S., Stoehr, G., Peplowska, K., Cox, J., Mann, M., and Storchova, Z. (2012). Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. Mol. Syst. Biol. *8*, 608.

Subtelny, A.O., Eichhorn, S.W., Chen, G.R., Sive, H., and Bartel, D.P. (2014). Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature *508*, 66–71.

Sury, M.D., McShane, E., Hernandez-Miranda, L.R., Birchmeier, C., and Selbach, M. (2015). Quantitative proteomics reveals dynamic interaction of c-Jun N-terminal kinase (JNK) with RNA transport granule proteins splicing factor proline- and glutamine-rich (Sfpq) and non-POU domain-containing octamer-binding protein (Nono) during neuronal differentiation. Mol. Cell. Proteomics *14*, 50–65.

tom Dieck, S., Kochen, L., Hanus, C., Heumüller, M., Bartnik, I., Nassim-Assir, B., Merk, K., Mosler, T., Garg, S., Bunse, S., et al. (2015). Direct visualization of newly synthesized target proteins in situ. Nat. Methods *12*, 411–414.

Toyama, B.H., Savas, J.N., Park, S.K., Harris, M.S., Ingolia, N.T., Yates, J.R., 3rd, and Hetzer, M.W. (2013). Identification of long-lived proteins reveals exceptional stability of essential cellular structures. Cell *154*, 971–982.

Tyler, R.E., Pearce, M.M.P., Shaler, T.A., Olzmann, J.A., Greenblatt, E.J., and Kopito, R.R. (2012). Unassembled CD147 is an endogenous endoplasmic reticulum-associated degradation substrate. Mol. Biol. Cell *23*, 4668–4678.

Vabulas, R.M., and Hartl, F.U. (2005). Protein synthesis upon acute nutrient restriction relies on proteasome function. Science *310*, 1960–1963.

Vavouri, T., Semple, J.I., Garcia-Verdugo, R., and Lehner, B. (2009). Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. Cell *138*, 198–208.

Ward, C.L., and Kopito, R.R. (1994). Intracellular turnover of cystic fibrosis transmembrane conductance regulator. Inefficient processing and rapid degradation of wild-type and mutant proteins. J. Biol. Chem. *269*, 25710–25718.

Wells, J.N., Bergendahl, L.T., and Marsh, J.A. (2016). Operon gene order is optimized for ordered protein complex assembly. Cell Rep. *14*, 679–685.

Wheatley, D.N., Giddings, M.R., and Inglis, M.S. (1980). Kinetics of degradation of "short-" and "long-lived" proteins in cultured mammalian cells. Cell Biol. Int. Rep. *4*, 1081–1090.

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G., McCoy, A., et al. (2011). Overview of the CCP4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. *67*, 235–242.

Wiśniewski, J.R., Zougman, A., and Mann, M. (2009). Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. J. Proteome Res. *8*, 5674–5678.

Xie, S.Q., Nie, P., Wang, Y., Wang, H., Li, H., Yang, Z., Liu, Y., Ren, J., and Xie, Z. (2016). RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling. Nucleic Acids Res. *44* (D1), D254–D258.

Current Biology
Magazine

CellPress

Correspondence

Evolution of condensin and cohesin complexes driven by replacement of Kite by Hawk proteins

Jonathan N. Wells[1],
Thomas G. Gligoris[2,*],
Kim A. Nasmyth[2],
and Joseph A. Marsh[1,*]

Mitotic chromosome condensation, sister chromatid cohesion, and higher order folding of interphase chromatin are mediated by condensin and cohesin, eukaryotic members of the SMC (structural maintenance of chromosomes)–kleisin protein family. Other members facilitate chromosome segregation in bacteria [1]. A hallmark of these complexes is the binding of the two ends of a kleisin subunit to the apices of V-shaped Smc dimers, creating a tripartite ring capable of entrapping DNA (Figure 1A). In addition to creating rings, kleisins recruit regulatory subunits. One family of regulators, namely Kite dimers (Kleisin interacting winged-helix tandem elements), interact with Smc–kleisin rings from bacteria, archaea and the eukaryotic Smc5-6 complex, but not with either condensin or cohesin [2]. These instead possess proteins containing HEAT (Huntingtin/EF3/PP2A/Tor1) repeat domains whose origin and distribution have not yet been characterized. Using a combination of profile Hidden Markov Model (HMM)-based homology searches, network analysis and structural alignments, we identify a common origin for these regulators, for which we propose the name Hawks, i.e. HEAT proteins associated with kleisins.

HEAT repeat proteins are a highly diverse family, a small subset of which regulate cohesin and condensins in eukaryotes (Figure 1A). Building on the recent description of the Kite family, we asked whether this subset descends from a common ancestral HEAT repeat protein. However, this question presents significant technical difficulties. Repetitive sequences can diverge rapidly; indeed, the average sequence identity between mammalian HEAT repeat proteins and insect orthologues is just ~13% [3]. This makes accurate sequence alignment challenging, and classical methods for homology detection fail on all but the most similar of these proteins. To tackle this problem we developed a novel network-based approach. Briefly, this utilises extensive profile–profile HMM searches [4] to generate a network, which is then clustered to reveal groups of paralogous proteins (for details, see Supplemental Methods and Figure S1A in Supplemental Information, published with this article online).

Applying this method to budding yeast, we find that amongst a large number of diverse clusters, all HEAT repeat proteins known to interact with α-, β- and γ- kleisins form a distinct cluster

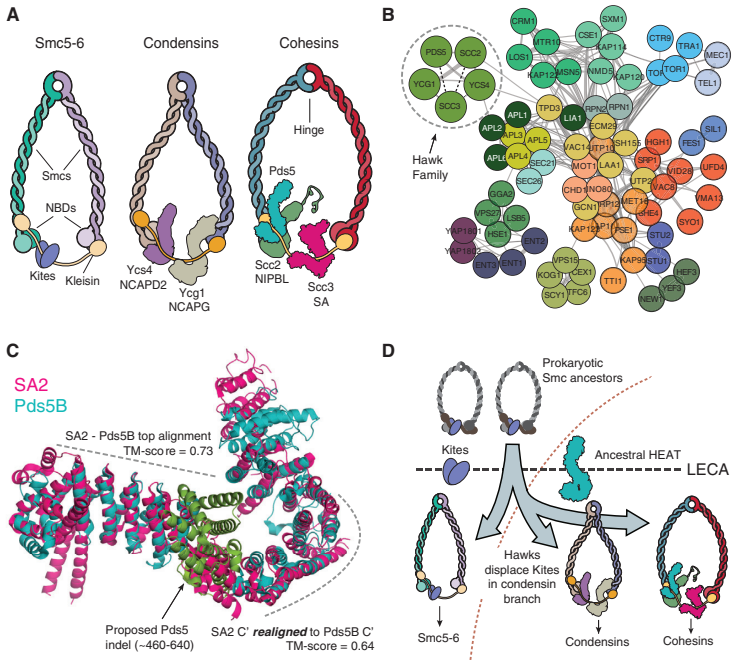


Figure 1. Hawk proteins form an evolutionarily related cluster and have displaced Kites in condensin and cohesin.
(A) The amino- and carboxy-terminal domains of the Smc polypeptides together form the globular nucleotide-binding domain (NBD). Kleisin subunits (yellow) then close the ring, topologically entrapping DNA in the process. In Smc5-6, Kite proteins interact with kleisins. In cohesin, Scc2 competes with Pds5 for its binding site on the kleisin. (B) Cohesin (Scc3, Scc2, Pds5) and condensin (Ycs4, Ycg1) HEAT regulators form a compact Hawk cluster (circled). Each cluster represented by a single colour. For clarity, only edges with a mean probability ≥99.0% are shown. Disconnected sub-graphs are hidden — the exception to this is Scc3, which is the weakest member of the hawk cluster and has been manually added (dashed edges). Above this threshold, the Hawk cluster has strong links to TPD3 (Protein phosphatase PP2A regulatory subunit A) and APL2 (Clathrin assembly protein large beta-1 chain). Members of the latter family (white labels) retain some of the strongest links to both hawks and lokiarchaeal proteins. (C) Despite a pairwise sequence identity of ~15%, SA2 and Pds5B (4PJU and 5HDT, respectively) are similar, with a TM-score of 0.44 (scores lower than 0.3 are spurious and alignment significance increases rapidly above 0.5). The top scoring local sequence alignment between the two HMM profiles was between Pds5B residues 311–418 and SA2 residues 285–400. Using TMalign to perform a pairwise structural alignment between these resulted in a fit with a TM-score of 0.73. The alignment is disrupted by a large indel in Pds5B — realigning SA2 to the region directly after this produces an improved TM-score of 0.64. The amino-terminal region of Pds5B has been truncated for clarity. (D) In the LECA Smc–kleisin ancestor, the Kite dimer was presumably flanked by the ancestral HEAT-protein/Hawk. Successive duplications of Hawks led to the Kites being displaced. The lack of Hawks in Smc5-6 suggests that it diverged earlier from the cohesins and condensins, whose specialised functions were facilitated by the recruitment of the Hawk family.

(Figure 1B). Similarly, in humans, two closely interacting clusters containing all condensin and cohesin Hawks are formed (Figure S1B,C and Figure S2A). These clusters are robust to changes in network parameters and are highly significant (p-value < 1 x 10^{-6}, permutation tests). Additionally, several other known protein families were recapitulated in individual clusters. GO-term analysis of biological processes also demonstrated highly significant enrichment in multiple clusters, e.g. the karyopherin α subunits, KPNA1–7. We therefore conclude that our method is effective and that Hawks form a distinct subgroup within the larger HEAT family.

Two important conclusions stem from these observations. First, NIPBL/Scc2 — the cohesin DNA loader — is confidently included in the Hawk cluster. This conclusion is now strongly supported by the recent biochemical studies of the *Chaetomium thermophilum* yeast Scc2, which is found to bind robustly to *C. thermophilum* Scc1 [5]. Second, our analysis fails to support the previous proposal that Nse5 and Nse6 associated with the eukaryotic Smc5-6 holocomplex contain HEAT repeats. Neither contains detectable repeats and searches for paralogues returned few proteins, none of which contained HEATs (see Supplemental Methods). Furthermore, a literature search revealed no evidence for Nse5 containing repeats, while the Nse6 annotation is based on a structural prediction which we were unable to replicate [6]. These negative findings indicate that Smc5-6 is alone amongst eukaryotic Smc–kleisin complexes in retaining Kites (the Nse1–3 subunits) and lacking Hawks.

We next turn to a possible origin for the Hawk family. Orthologues were found in almost all eukaryote species we tested, collectively accounting for all major extant branches of the eukaryotic tree. We searched for related sequences in Lokiarchaeota, currently the closest known archaeal relatives of the Last Eukaryotic Common Ancestor (LECA). Several lokiarchaeal HEAT repeat proteins produced significant alignments with Hawks; furthermore, we find that the lokiarchaeal HEATs predominantly cluster with clathrin adaptor proteins, which share sequence and structural similarity with the Hawks (Figure 1B and Figure S2B) [7]) These observations lead us to tentatively suggest that the ancestral Hawk protein derived from an ancient group of HEAT proteins related to the clathrin adaptor family, and that this occurred close to or even prior to the prokaryote–eukaryote split.

An independent test of our conclusion that Hawks derive from a common ancestor is provided by structural analysis of yeast subunits Pds5/Pds5B and Scc3/SA2 (e.g. [8,9]). From sequence analysis we see that the remaining Hawks are of a similar size, with similar distributions of repeats identified from sequence, particularly when compared to the clathrin adaptors (Figure S2C). Pds5B and SA2 also align well structurally, though disrupted by an indel in Pds5B (Figure 1C and Figure S2D) [10]. When this region was omitted, the alignment improved considerably (Figure 1C). Finally, SA2 and Pds5B display similar patterns of conservation along their spines (Figure S2E). These similarities between SA2 and Pds5B are particularly striking since SA2/Scc3 appears to be the most diverged member of the Hawk clusters. Supporting this further, the structure of Scc2 from *C. thermophilum* [5] shows that its carboxy-terminal region has a very similar shape and structure to Pds5, albeit lacking the indel found in the latter.

Based on our main conclusions — the strong clustering of Hawks, their deep conservation across eukaryotes, and their absence from Smc5-6 complexes — we propose a model for the Smc-kleisin complex in LECA (Figure 1D). According to our hypothesis, the ancestral Hawk protein was recruited to the complex very early in eukaryotic history. Successive duplications of this protein displaced the Kites, leading to the predecessor of modern condensins, containing two Hawk regulators (similar to budding yeast's Ycs4 and Ycg1), and to the cohesins, with three (Pds5, Scc3 and Scc2). An outstanding question from this model is whether or not Smc5-6 gained and then lost Hawks, or whether its lack of Hawks indicates that it forms a branch distinct from the cohesins and condensins. In any case, our results show that the Smc–kleisins can be separated into two groups — those containing Kites, and those containing Hawks; of these, the Hawk–Smc–kleisins appear to be uniquely eukaryotic. Both cohesin and condensin share the ability to organize loops of chromatin fibres around an axial core and cohesin may have later acquired the ability to hold sister chromatids together and to be cleaved by separase. A key question now emerging is whether the replacement of Kites by Hawks was mechanistically associated with the acquisition of these novel functions.

SUPPLEMENTAL INFORMATION

Supplemental Information contains methods and two figures, and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2016.11.050.

ACKNOWLEDGEMENTS

We wish to thank Hongtao Yu for kindly allowing us to discuss in earlier versions of this manuscript unpublished Scc2 data. Research in J.A.M. lab is supported by an MRC Career Development Award (MR/M02122X/1) and in the K.A.N lab by an MRC Research Grant (MR/L018047/1).

REFERENCES

1. Nasmyth, K., and Haering, C.H. (2005). The structure and function of SMC and kleisin complexes. Annu. Rev. Biochem. *74*, 595–648.
2. Palecek, J.J., and Gruber, S. (2015). Kite proteins: a superfamily of SMC/Kleisin partners conserved across Bacteria, Archaea, and Eukaryotes. Structure *23*, 2183–2190.
3. Andrade, M.A., Petosa, C., O'Donoghue, S.I., Müller, C.W., and Bork, P. (2001). Comparison of ARM and HEAT protein repeats. J. Mol. Biol. *309*, 1–18.
4. Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods *9*, 173–175.
5. Kikuchi, S., Borek, D.M., Otwinowski, Z., Tomchick, D.R., and Yu, H. (2016). Crystal structure of the cohesin loader Scc2 and insight into cohesinopathy. Proc. Natl. Acad. Sci. USA *113*, 201611333.
6. Pebernard, S., Wohlschlegel, J., McDonald, W.H., Yates, J.R., and Boddy, M.N. (2006). The Nse5-Nse6 dimer mediates DNA repair roles of the Smc5-Smc6 complex. Mol. Cell Biol. *26*, 1617–1630.
7. Neuwald, A.F., and Hirano, T. (2000). HEAT repeats associated with condensins, cohesins, and other complexes involved in chromosome-related functions. Genome Res.*10*, 1445–1452.
8. Hara, K., Zheng, G., Qu, Q., Liu, H., Ouyang, Z., Chen, Z., Tomchick, D.R., and Yu, H. (2014). Structure of cohesin subcomplex pinpoints direct shugoshin-Wapl antagonism in centromeric cohesion. Nat. Struct. Mol. Biol. *21*, 864–870.
9. Ouyang, Z., Zheng, G., Tomchick, D.R., Luo, X., and Yu, H. (2016). Structural basis and IP6 requirement for Pds5-dependent cohesin dynamics. Mol. Cell *62*, 248–259.
10. Lee, B.-G., Roig, M.B., Jansma, M., Petela, N., Metson, J., Nasmyth, K., and Löwe, J. (2016). Crystal structure of the cohesin gatekeeper Pds5 and in complex with kleisin Scc1. Cell Rep. *14*, 2108–2115.

[1]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK.
[2]Department of Biochemistry, University of Oxford, Oxford, OX1 3QU, UK.
*E-mail: thomas.gligoris@bioch.ox.ac.uk (T.G.G.), joseph.marsh@igmm.ed.ac.uk (J.A.M.)

Co-translational assembly of protein complexes

Jonathan N. Wells*, L. Therese Bergendahl* and Joseph A. Marsh*[1]
*MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, U.K.

Abstract

The interaction of biological macromolecules is a fundamental attribute of cellular life. Proteins, in particular, often form stable complexes with one another. Although the importance of protein complexes is widely recognized, we still have only a very limited understanding of the mechanisms underlying their assembly within cells. In this article, we review the available evidence for one such mechanism, namely the coupling of protein complex assembly to translation at the polysome. We discuss research showing that co-translational assembly can occur in both prokaryotic and eukaryotic organisms and can have important implications for the correct functioning of the complexes that result. Co-translational assembly can occur for both homomeric and heteromeric protein complexes and for both proteins that are translated directly into the cytoplasm and those that are translated into or across membranes. Finally, we discuss the properties of proteins that are most likely to be associated with co-translational assembly.

Introduction

Proteins within the cell must carry out their important functions in an environment that is highly crowded and are in constant physical contact with various other proteins, metabolites and macromolecules [1,2]. Apart from many transient interactions, which may or may not have important functional duties, e.g. cellular signalling [3,4], many if not most intracellular proteins function as subunits of more long-lived protein complexes [5,6]. Despite the fact that complex formation is crucial for understanding the function (and malfunction) of many proteins, the fundamental mechanisms behind the assembly of individual proteins into complexes with defined quaternary structure have often been neglected and little is known about the *in vivo* assembly process. One interesting aspect of protein complex assembly is the degree to which it is coupled to the cellular translation machinery. Are the subunits fully translated before finding their interaction partners and forming their defined quaternary structure in a post-translational assembly pathway? Or do some protein interactions form as the individual subunits are still being translated, through co-translational assembly?

The general process of the maturation of a functioning protein involves folding and translocation of the polypeptide as well as complex assembly. In recent years, there has been a growing body of evidence showing that the protein folding process can often occur while the polypeptide is being translated, i.e. co-translationally involving the nascent polypeptide chain [7–12]. Although this could arise due to the basic energetics of protein folding, there are also potential functional benefits to co-translational folding. For example, it may provide a means of tuning the potential energy landscape by lowering the energy of folding intermediates. A co-translational folding process also contributes to the earlier formation of secondary and tertiary structures making unfavourable inter-domain aggregation events less likely.

The assembly of protein complexes can, in many ways, be considered analogous to protein folding, in that it typically follows a specific pathway via energetically favourable assembly intermediates [13,14]. It is therefore natural to envisage that assembly could also sometimes occur co-translationally and even be functionally advantageous. Many protein complex subunits are highly dynamic or unstable in isolation [15] and so rapid assembly during translation minimizes the opportunities for misfolding or aggregation. If the subunits assemble through a series of lower energy intermediates of nascent polypeptides, it would lower the overall energy of the assembly, analogous to the tuning of folding. Co-translational protein interactions can also be viewed as a means of ensuring a precisely ordered assembly process and for avoiding unfavourable inter-subunit aggregation. Contrary to co-translational folding however, the kinetics of co-translational assembly are not only a function of the rate of assembly, but also highly dependent on the concentration of available assembly partners in close proximity to the polysome.

Although co-translational assembly has received much less attention than co-translational folding, numerous cases have been reported over the years, beginning with observations of specific proteins associating co-translationally with the cytoskeleton [16–19]. Furthermore, recent evidence suggests that the phenomenon might be widespread [20]. Here we review several examples of co-translational assembly, discussing reasons why it occurs and the functional benefits that it can provide. Although there are many transient protein–protein interactions that occur co-translationally, e.g. involving chaperones [21] or targeting signal sequences for

Key words: co-translational assembly, heteromer, homomer, protein interactions, quaternary structure, translation.

Abbreviations: COMPASS, complex proteins associated with Set1p; NF-κB1, nuclear factor of kappa light polypeptide gene enhancer in B-cells 1; RHD, Rel homology domain; ER, endoplasmic reticulum; hERG1, human ether-à-go-go related gene.

[1] **To whom correspondence should be addressed (email joseph.marsh@igmm.ed.ac.uk).**

translocation [22], here we will focus on the assembly of stable protein complexes.

Co-translational assembly of homomers

Protein complexes can be broadly split into two categories based upon their quaternary structure: homomers, formed from the self-assembly of a single subunit type and heteromers, formed from multiple distinct subunits. There are two ways in which co-translational homomer assembly can occur. In one, a newly translated subunit is released and then interacts with another still-translating copy of itself, most probably on the same polysome from which it was translated (Figure 1A). Alternatively, co-translational assembly could involve interaction between two nascent chains on the same polysome (Figure 1B).

An early example of co-translational assembly of a homomer came from investigations into the well-characterized bacterial homotetramer β-galactosidase [23,24]. The enzymatic activity of β-galactosidase is only evident after the conformational changes that are required for tetramer formation have taken place. In these experiments, it was shown that the enzymatically active form of the complex could be observed at the same time as nascent polypeptide chains. It became clear that not only the folding, but also the assembly into the functioning enzyme occurred in a co-translational manner. The authors suggested that this was due to the proximity of the nascent polypeptides, as monomers from adjacent ribosomes dimerized before forming the final tetrameric structure.

The reovirus attachment protein σ1 forms a homotrimer and can be divided into two segments: an N-terminal tail that is anchored in the virion and a globular C-terminal domain that is responsible for virion attachment (Figure 2). Curiously it was found that the trimerization of the two regions takes place using two different mechanisms [25]. Assembly of the N-terminal region, which is translated first, was found to occur co-translationally, at neighbouring ribosomes that had passed the midpoint of the mRNA strand. In contrast, trimerization of the globular C-terminal region, which is translated last, was found to be highly chaperone- and ATP-dependent and occurs post-translationally.

The tumour suppressor p53 forms a homotetramer with dihedral symmetry. Although both alleles of p53 are often mutated or non-functional in cancer cells, mutations in a single allele often display a dominant-negative effect. Depending on the location of the mutation, numerous factors contribute to this effect, but a key aspect relates to the mechanism of p53 tetramer assembly and the extent to which it is coupled to translation. It was demonstrated that this process occurs by an initial co-translational dimerization of p53, with tetramers forming separately and post-translationally [26]. The suggested driving force for this assembly mechanism was the stabilization of the dimer through hydrophobic interactions between the N-termini. A direct effect of this co-translational assembly is that the possible stoichiometries of the fully assembled complex

Figure 1 | Co-translational assembly of protein complexes

In all panels, moving from left (5′) to right (3′) on the polysome (i.e. the mRNA bound to multiple ribosomes), we can see increasingly long nascent chains being translated. Homomer assembly can occur in two ways. In (**A**), a full-length subunit is released and binds to a nascent chain, forming a co-translationally assembled homodimer. In (**B**), two nascent chains from the same polysome interact with each other. For heteromer assembly (**C**), a different subunit (red) encoded by a different gene binds to a nascent chain, forming a co-translationally assembled heterodimer. These are hypothetical examples of co-translational assembly based upon PDB ID: 2I99 (homodimer) and PDB ID: 2DCU (heterodimer).

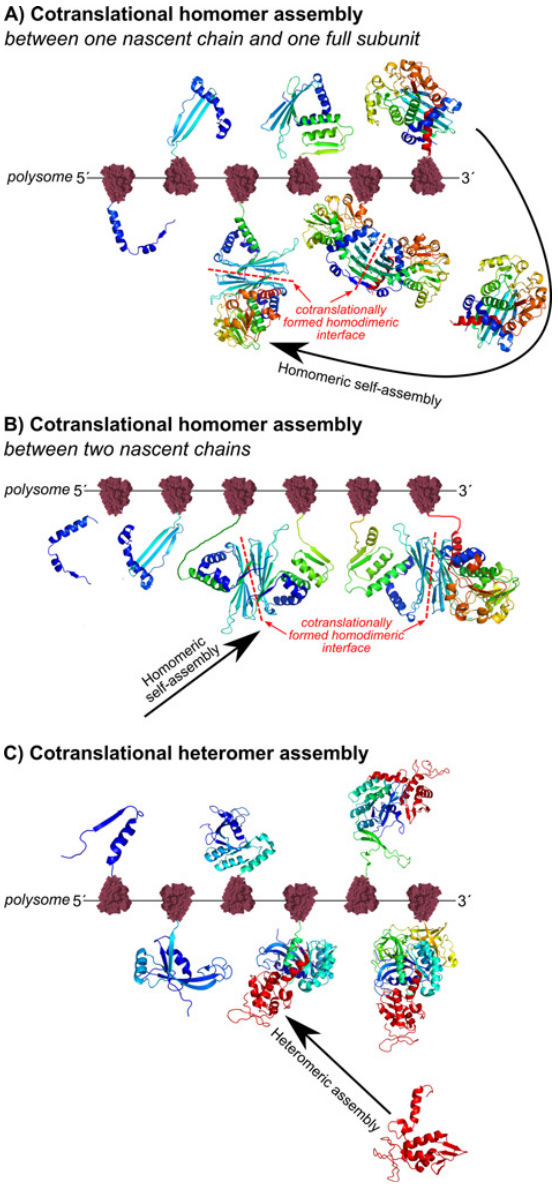
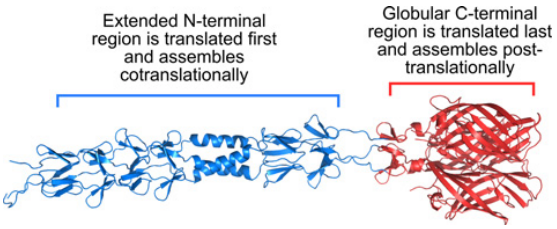


Figure 2 | Structure of the reovirus attachment protein σ1

In this homotrimeric complex (PDB ID: 3S6X), the extended N-terminal region (blue) is known to assemble co-translationally, whereas the globular C-terminal region (red) assembles only post-translationally. This highlights the idea that co-translationally forming interfaces should generally localized towards the N-termini of proteins, as they will spend more time as part of a nascent chain and have more time to co-translationally interact.



are constrained: p53 dimers will always be homomers of either the wild-type or the mutant version of the protein. Consequently, one-fourth of the resulting homotetramers will be fully wild-type, as opposed to only one-sixteenth in a situation where co-translational dimerization does not occur. This suggests that the co-translational dimerization step has a strong influence on the magnitude of the dominant-negative effect observed.

NF-κB1 (nuclear factor of kappa light polypeptide gene enhancer in B-cells 1) is a member of the NF-κB family of transcription factors and is involved in regulation of several cellular processes, particularly the inflammatory response [27]. The complex exists predominantly as a heterodimer of p50 and p105 subunits, with the p50 subunit being a truncated form of p105. Full-length p105 comprises an N-terminal Rel homology domain (RHD) and a larger C-terminal ankyrin-repeat domain that functions as an I-κB-like inhibitor of mature NF-κB1. Between these two domains lie a nuclear-localization signal and a glycine-rich region that acts as the site of endoproteolytic cleavage by the 26S proteasome [28]. Active NF-κB1 requires cleavage and degradation of the C-terminal domain of p105 to form mature p50 [29]. A key question arising from this observation is how the proteasome degrades p105 while sparing p50. Building on early observations that free p50 rapidly associates with other Rel family proteins *in vitro*, Lin et al. [30] demonstrated *in vivo* that p50–p105 heterodimers assemble on the same polysome via co-translational homodimerization of p50 RHDs [31] (Figure 3). This is coupled with co-translational processing by the proteasome; crucially, it is the process of dimerization that appears to act as a physical barrier to degradation of p50. In support of this, it was shown that deletion of the second sub-domain of RHD (essential for dimerization) led to a significant reduction in the amount of p50 observed upon expression of the mutant *NF-κB1* gene. This suggests that in the absence of dimerization, p105 is completely degraded. If so, this provides a clear example of how co-translational assembly can be functionally important; in this case, co-translational assembly is essential for the production of mature NF-κB1, with the active p50 subunit being placed under immediate control of the inhibitory p105 subunit by the process. Subsequent post-translational activation then depends on phosphorylation and ubiquitin-mediated cleavage of the remaining p105 ankyrin-repeat domain.

Co-translational assembly of heteromers

The assembly of heteromers is inherently more complex than for homomers, due to the fact that it involves interactions between distinct proteins that are usually encoded by different genes. Those interacting proteins must somehow find each other within the cell. Co-translational interaction, in which a fully translated protein finds its way to the nascent chain of another protein (Figure 1C), provides a way to minimize the stochasticity of assembly by increasing the chance of subunit encounter.

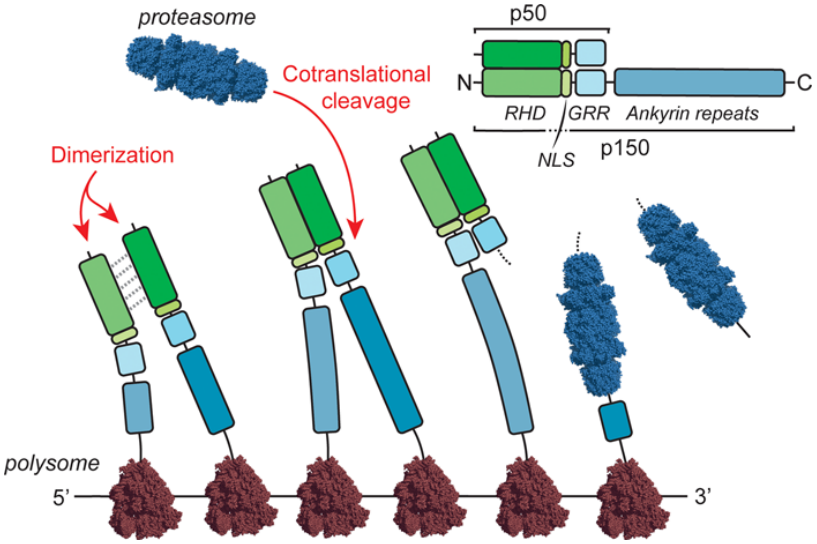
One example of a heteromeric interaction with both co- and post-translational assembly mechanisms is the covalent disulfide bond formation between heavy and light chains in the immunoglobulin molecule. Despite earlier evidence of post-translational formation of the disulfide linkers, it was shown that over-production of light chains in the endoplasmic reticulum (ER) in certain cell types lead to light-heavy chain heterodimerization on the nascent heavy chain, purely due to light chain abundance in the proximity of the translating heavy chain transcript [32]. Thus protein expression levels and abundance are likely to be important regulators of whether or not heteromeric assembly occurs co-translationally.

The yeast histone methyltransferase, COMPASS (complex proteins associated with Set1p), is comprised of eight different subunits. In work originally designed to investigate the role of mRNA in COMPASS function, a four-member sub-complex of COMPASS (formed by the proteins Swd1p, Spp1p, Shg1p and Set1p) was found to interact with *SET1* mRNA [33]. Crucially, formation of the mRNA-associated sub-complex was found to be dependent on active translation, indicating that the subunits are binding to the nascent Set1p protein as it is translated. Furthermore, whereas structural data are not available for the full complex, it appears that the binding sites of Shg1p, Swd1p and Spp1p are localized to the N-terminal or central region of Set1p; this is consistent with co-translational assembly as these regions are translated earlier [33,34].

Previously, the first systematic analysis of co-translational assembly has been performed. Duncan and Mata [20] used ribonucleoprotein immunoprecipitation-microarray (RIP-Chip) experiments to identify mRNA sequences associated with 31 proteins from *Schizosaccharomyces pombe* [20]. Here they found that 12 of these (38%) co-purified with the mRNAs of known interaction partners. Importantly, as for the COMPASS example, co-purification was found to be dependent on active translation, indicating that interactions are probably occurring between proteins and nascent peptides, rather than protein and mRNA. These

Figure 3 | **Co-translational assembly of p50–p105 heterodimer**

The p50 protein is ∼400 residues in length and comprises of a RHD, nuclear localization signal (NLS) and C-terminal glycine-rich region (GRR), which is targeted by the proteasome. The p105 protein differs only in that it contains an additional ankyrin-repeat domain. During translation, the RHDs of two nascent polypeptides dimerize, though it is unclear as to whether this occurs while both chains are being actively translated (as shown here) or between freshly synthesized p105 and the actively translating chain, as in Figure 1(A). As very rapid dimerization of the RHDs is essential to prevent complete degradation of p50–p105 by the proteasome (which also occurs co-translationally), it seems plausible that the former scenario is correct.



interactions were also found to be highly specific; Cdc2p, for example, was found to co-purify with just two mRNAs, despite having a large number of known and hypothesized protein interaction partners. Interestingly, the fact that these mRNA–protein interactions are so specific has since been used by the same group to predict novel protein–protein interactions [35].

Co-translational assembly of secreted and membrane complexes

The above examples involve co-translational assembly within the cytoplasm, but many proteins are directly translocated into or across membranes during translation. In eukaryotes, membrane and secreted proteins are translated at the rough ER. In investigations into the assembly of the extracellular human tenascin protein, responsible for cell adhesion, it was shown that the hexameric complex is formed without any assembly intermediates being observed [36]. As soon as the tenascin protein is experimentally detectable, it appears to be co-translationally assembled into its active hexamer structure. In this case, the authors suggested that the arrangement of the membrane-bound polysome at the ER, where the ribosomes have been seen to form various circular loops and spirals, directly resulted in the homomer acquiring its circular hexamer shape.

A further example of ER membrane influence on co-translational assembly is seen in voltage-gated potassium cation channels. These channels are tetrameric with interfaces located at the N-terminal region of the subunits [referred to as the tetramerization (T1)-domains]. In experiments using *Xenopus laevis* oocytes, it was shown that T1–T1 association occurred between ribosome-attached subunits and the ER membrane was postulated to regulate the local concentration of the interacting domains [37]. In the related human ether-à-go-go related gene (*hERG1*), responsible for the potassium channel hERG, the two subunits hERG1a and hERG1b are isoforms of *hERG1*, arising from two mRNA splice-variants [38]. The isoforms are identical apart from the important N-termini and it was observed that the two N-termini localize and bind to each other co-translationally. This mechanism is crucial to avoid unfavourable aggregation events involving the hERG1b subunits and is mediated by the ER, which ensures co-localization of the transcripts.

Finally, there is evidence that the plant D1 transmembrane protein assembles co-translationally into the photosystem II complex [39,40]. This is an interesting example as the D1 protein frequently experiences photodamage and experiences a high rate of turnover. Thus the ability of D1 to be translated directly into the chloroplast membrane and co-translationally assemble allows photosystem II to be quickly repaired. It is also notable that translational pausing is known to occur at

specific sites during the translation of D1 [41], potentially allowing time for assembly to occur [42], analogous to how translational pausing can facilitate protein folding [12].

Perspectives

Here we have highlighted a number of examples of homomeric and heteromeric protein complexes that assemble co-translationally. However, we still have very little idea about the frequency of the phenomenon. One systematic analysis suggested that it might be quite widespread, yet this work considered only a very small fraction of known proteins in fission yeast [20]. In addition, questions remain about the specific mechanisms by which co-translational interactions occur. For example, it is unclear whether binding events are limited to those occurring between one nascent and one fully-folded chain or whether dimerization ever occur while both chains are being translated. Thus, there is considerable future potential for both large- and small-scale screens looking for evidence of co-translational assembly.

Why do some protein complexes assemble co-translationally whereas others do not? Although possible functional benefits have been discussed here, it is important to remember that co-translational assembly has not necessarily been selected for evolutionarily in all cases. Co-translational assembly could occur simply because a free subunit encounters a nascent chain and their interaction is energetically favourable. In fact, for some proteins, there may be evolutionary pressure to avoid co-translational assembly. Although we only have experimental evidence of co-translational assembly for a fairly small number of complexes, we can make some predictions about which complexes might be most likely to assemble co-translationally:

- All things being equal, homomers should be more likely to co-translationally assemble than heteromers, since interacting subunits can be translated from the same polysome and local subunit concentration will be high.
- Many prokaryotic complexes are encoded in operons, so that interacting proteins are often translated off the same polycistronic mRNA. This ensures that interacting subunits are in close physical proximity upon translation and facilitates a higher rate of complex assembly [43]. Thus we can predict that operon-encoded heteromers should be more likely to undergo co-translational assembly.
- For both homomers and heteromers, the likelihood of co-translational assembly should be greater for highly abundant proteins, as this will increase the chance that an interaction partner encounters and binds a nascent chain still in the process of being translated.
- Localization towards N-terminal regions is likely to be a general feature of interfaces that form co-translationally, since this will allow more time for co-translational assembly to occur. Therefore, complexes with N-terminal interfaces should be more likely to have formed co-translationally.
- Subunits that are highly flexible or disordered in isolation [15] could benefit from co-translational assembly, as this would avoid them spending unnecessary time free and susceptible to proteases in the cell.
- The first step of a protein complex assembly pathway is the most probable to occur co-translationally. Thus we may be able to use experimental characterization or structure-based prediction of assembly order [13,14] to identify subunits and interfaces that are most likely to form co-translationally.

Finally, there are major questions remaining about how co-translational assembly is regulated and how proteins are localized to polysome, especially for heteromers with subunits translated from different mRNA molecules. This is especially important for eukaryotic complexes, which have a much greater propensity to form heteromers [44,45], compared with bacterial proteins, which are more likely to self-assemble into homomers [46] or be encoded in operons. Much more work is needed to fully understand how the assembly of heteromeric complexes occurs within eukaryotic cells, both co- and post-translationally and how it is regulated.

Acknowledgements

We thank Cathy Abbott and Dinesh Soares for helpful comments on the manuscript.

Funding

This work was supported by a University of Edinburgh Chancellor's Fellowship to J.M.

References

1 Ellis, R.J. (2001) Macromolecular crowding: an important but neglected aspect of the intracellular environment. Curr. Opin. Struct. Biol. **11**, 114–119 CrossRef PubMed
2 McGuffee, S.R. and Elcock, A.H. (2010) Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. PLoS Comput. Biol. **6**, e1000694 CrossRef PubMed
3 Nooren, I.M.A. and Thornton, J.M. (2003) Structural characterisation and functional significance of transient protein-protein interactions. J. Mol. Biol. **325**, 991–1018 CrossRef PubMed
4 Landry, C.R., Levy, E.D., Abd Rabbo, D., Tarassov, K. and Michnick, S.W. (2013) Extracting insight from noisy cellular networks. Cell **155**, 983–989 CrossRef PubMed
5 Perica, T., Marsh, J.A., Sousa, F.L., Natan, E., Colwell, L.J., Ahnert, S.E. and Teichmann, S.A. (2012) The emergence of protein complexes: quaternary structure, dynamics and allostery. Biochem. Soc. Trans. **40**, 475–491 CrossRef PubMed
6 Marsh, J.A. and Teichmann, S.A. (2014) Structure, dynamics, assembly, and evolution of protein complexes. Annu. Rev. Biochem. **84**, 551–575 CrossRef PubMed
7 Komar, A.A. (2009) A pause for thought along the co-translational folding pathway. Trends Biochem. Sci. **34**, 16–24 CrossRef PubMed
8 O'Brien, E.P., Vendruscolo, M. and Dobson, C.M. (2012) Prediction of variable translation rate effects on cotranslational protein folding. Nat. Commun. **3**, 868 CrossRef PubMed
9 Gloge, F., Becker, A.H., Kramer, G. and Bukau, B. (2014) Co-translational mechanisms of protein maturation. Curr. Opin. Struct. Biol. **24**, 24–33 CrossRef PubMed

10 Ciryam, P., Morimoto, R.I., Vendruscolo, M., Dobson, C.M. and O'Brien, E.P. (2013) *In vivo* translation rates can substantially delay the cotranslational folding of the *Escherichia coli* cytosolic proteome. Proc. Natl. Acad. Sci. U.S.A. **110**, E132–E140 CrossRef PubMed

11 Pechmann, S. and Frydman, J. (2013) Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat. Struct. Mol. Biol. **20**, 237–243 CrossRef PubMed

12 Zhang, G., Hubalewska, M. and Ignatova, Z. (2009) Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat. Struct. Mol. Biol. **16**, 274–280 CrossRef PubMed

13 Levy, E.D., Boeri Erba, E., Robinson, C.V. and Teichmann, S.A. (2008) Assembly reflects evolution of protein complexes. Nature **453**, 1262–1265 CrossRef PubMed

14 Marsh, J.A., Hernández, H., Hall, Z., Ahnert, S.E., Perica, T., Robinson, C.V. and Teichmann, S.A. (2013) Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell **153**, 461–470 CrossRef PubMed

15 Marsh, J.A., Teichmann, S.A. and Forman-Kay, J.D. (2012) Probing the diverse landscape of protein flexibility and binding. Curr. Opin. Struct. Biol. **22**, 643–650 CrossRef PubMed

16 Isaacs, W.B., Kim, I.S., Struve, a. and Fulton, A.B. (1989) Biosynthesis of titin in cultured skeletal muscle cells. J. Cell Biol. **109**, 2189–2195 CrossRef PubMed

17 Isaacs, W.B. and Fulton, A.B. (1987) Cotranslational assembly of myosin heavy chain in developing cultured skeletal muscle. Proc. Natl. Acad. Sci. U.S.A. **84**, 6174–6178 CrossRef PubMed

18 L'Ecuyer, T.J., Noller, J.A. and Fulton, A.B. (1998) Assembly of tropomyosin isoforms into the cytoskeleton of avian muscle cells. Pediatr. Res. **43**, 813–822 CrossRef PubMed

19 Isaacs, W.B., Cook, R.K., Van Atta, J.C., Redmond, C.M. and Fulton, A.B. (1989) Assembly of vimentin in cultured cells varies with cell type. J. Biol. Chem. **264**, 17953–17960 PubMed

20 Duncan, C.D. and Mata, J. (2011) Widespread cotranslational formation of protein complexes. PLoS Genet. **7**, e1002398 CrossRef PubMed

21 Zhang, Y., Ma, C., Yuan, Y., Zhu, J., Li, N., Chen, C., Wu, S., Yu, L., Lei, J. and Gao, N. (2014) Structural basis for interaction of a cotranslational chaperone with the eukaryotic ribosome. Nat. Struct. Mol. Biol. **21**, 1042–1046 CrossRef PubMed

22 Park, E. and Rapoport, T.A. (2012) Mechanisms of Sec 61/SecY-mediated protein translocation across membranes. Annu. Rev. Biophys. **41**, 21–40 CrossRef PubMed

23 Zipser, D. and Perrin, D. (1963) Complementation on ribosomes. Cold Spring Harb. Symp. Quant. Biol. **28**, 533–537 CrossRef

24 Kiho, Y. and Rich, A. (1964) Induced enzyme formed on bacterial polyribosomes. Proc. Natl. Acad. Sci. U.S.A. **51**, 111–118 CrossRef PubMed

25 Leone, G., Coffey, M.C., Gilmore, R., Duncan, R., Maybaum, L. and Lee, P.W.K. (1996) C-terminal trimerization, but not N-terminal trimerization, of the reovirus cell attachment protein is a posttranslational and Hsp70/ATP-dependent process. J. Biol. Chem. **271**, 8466–8471 CrossRef PubMed

26 Nicholls, C.D., McLure, K.G., Shields, M.A. and Lee, P.W.K. (2002) Biogenesis of p53 Involves cotranslational dimerization of monomers and posttranslational dimerization of dimers: implications on the dominant negative effect. J. Biol. Chem. **277**, 12937–12945 CrossRef PubMed

27 Jurk, D., Wilson, C., Passos, J.F., Oakley, F., Correia-Melo, C., Greaves, L., Saretzki, G., Fox, C., Lawless, C., Anderson, R. et al. (2014) Chronic inflammation induces telomere dysfunction and accelerates ageing in mice. Nat. Commun. **2**, 4172 CrossRef PubMed

28 Lin, L. and Ghosh, S. (1996) A glycine-rich region in NF-kappaB p105 functions as a processing signal for the generation of the p50 subunit. Mol. Cell. Biol. **16**, 2248–2254 PubMed

29 Fan, C.M. and Maniatis, T. (1991) Generation of p50 subunit of NF-kappa B by processing of p105 through an ATP-dependent pathway. Nature **354**, 395–398 CrossRef PubMed

30 Lin, L., DeMartino, G.N. and Greene, W.C. (2000) Cotranslational dimerization of the rel homology domain of NF-kappaB1 generates p50-p105 heterodimers and is required for effective p50 production. EMBO J. **19**, 4712–4722 CrossRef PubMed

31 Chen, F.E., Huang, D.B., Chen, Y.Q. and Ghosh, G. (1998) Crystal structure of p50/p65 heterodimer of transcription factor NF-kappaB bound to DNA. Nature **391**, 410–413 CrossRef PubMed

32 Bergman, L.W. and Kuehl, W.M. (1979) Formation of intermolecular disulfide bonds on nascent immunoglobulin polypeptides. J. Biol. Chem. **254**, 5690–5694 PubMed

33 Halbach, A., Zhang, H., Wengi, A., Jablonska, Z., Gruber, I.M.L., Halbeisen, R.E., Dehé, P.-M., Kemmeren, P., Holstege, F., Géli, V. et al. (2009) Cotranslational assembly of the yeast SET1C histone methyltransferase complex. EMBO J. **28**, 2959–2970 CrossRef PubMed

34 Dehé, P.-M. and Géli, V. (2006) The multiple faces of set1. Biochem. Cell Biol. **84**, 536–548 CrossRef PubMed

35 Duncan, C.D. and Mata, J. (2014) Cotranslational protein-RNA associations predict protein-protein interactions. BMC Genomics **15**, 298 CrossRef PubMed

36 Redick, S.D. and Schwarzbauer, J.E. (1995) Rapid intracellular assembly of tenascin hexabrachions suggests a novel cotranslational process. J. Cell Sci. **108** (Pt 4), 1761–1769 PubMed

37 Lu, J., Robinson, J.M., Edwards, D. and Deutsch, C. (2001) T1-T1 interactions occur in ER membranes while nascent KV peptides are still attached to ribosomes. Biochemistry **40**, 10934–10946 CrossRef PubMed

38 Phartiyal, P., Jones, E.M.C. and Robertson, G.A. (2007) Heteromeric assembly of human ether-à-go-go-related gene (hERG) 1a/1b channels occurs cotranslationally via N-terminal interactions. J. Biol. Chem. **282**, 9874–9882 CrossRef PubMed

39 Zhang, L., Paakkarinen, V., Van Wijk, K.J. and Aro, E.M. (1999) Co-translational assembly of the D1 protein into photosystem II. J. Biol. Chem. **274**, 16062–16067 CrossRef PubMed

40 Zhang, L. and Aro, E.M. (2002) Synthesis, membrane insertion and assembly of the chloroplast-encoded D1 protein into photosystem II. FEBS Lett. **512**, 13–18 CrossRef PubMed

41 Kim, J., Klein, P.G. and Mullet, J.E. (1991) Ribosomes pause at specific sites during synthesis of membrane-bound chloroplast reaction center protein D1. J. Biol. Chem. **266**, 14931–14938 PubMed

42 Képès, F. (1996) The "+70 pause": hypothesis of a translational control of membrane protein assembly. J. Mol. Biol. **262**, 77–86 CrossRef PubMed

43 Sneppen, K., Pedersen, S., Krishna, S., Dodd, I. and Semsey, S. (2010) Economy of operon formation: cotranscription minimizes shortfall in protein complexes. MBio. **1**, 3–5 CrossRef

44 Lynch, M. (2012) The evolution of multimeric protein assemblages. Mol. Biol. Evol. **29**, 1353–1366 CrossRef PubMed

45 Marsh, J.A. and Teichmann, S.A. (2014) Protein flexibility facilitates quaternary structure assembly and evolution. PLoS Biol. **12**, e1001870 CrossRef PubMed

46 Marsh, J.A., Rees, H.A., Ahnert, S.E. and Teichmann, S.A. (2015) Structural and evolutionary versatility in protein complexes with uneven stoichiometry. Nat. Commun. **6**, 6394 CrossRef PubMed

Received 17 July 2015
doi:10.1042/BST20150159

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Regulation, evolution and consequences of cotranslational protein complex assembly

Eviatar Natan[1], Jonathan N Wells[2], Sarah A Teichmann[3] and Joseph A Marsh[2]

Most proteins assemble into complexes, which are involved in almost all cellular processes. Thus it is crucial for cell viability that mechanisms for correct assembly exist. The timing of assembly plays a key role in determining the fate of the protein: if the protein is allowed to diffuse into the crowded cellular milieu, it runs the risk of forming non-specific interactions, potentially leading to aggregation or other deleterious outcomes. It is therefore expected that strong regulatory mechanisms should exist to ensure efficient assembly. In this review we discuss the cotranslational assembly of protein complexes and discuss how it occurs, ways in which it is regulated, potential disadvantages of cotranslational interactions between proteins and the implications for the inheritance of dominant-negative genetic disorders.

Addresses
[1] Department of Chemistry, University of Oxford, 12 Mansfield Rd, Oxford OX1 3TA, UK
[2] MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK
[3] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Corresponding authors: Natan, Eviatar (eviatarhj@gmail.com) and Marsh, Joseph A (joseph.marsh@igmm.ed.ac.uk)

Current Opinion in Structural Biology 2017, **42**:90–97

This review comes from a themed issue on **Folding and binding**

Edited by **Jane Clarke** and **Rohit V Pappu**

For a complete overview see the Issue and the Editorial

Available online 12th December 2016

http://dx.doi.org/10.1016/j.sbi.2016.11.023

Introduction

Many proteins can assemble into protein complexes [1,2•]. Although there is tremendous diversity in the types of quaternary structures that can be formed [3,4••], at the simplest level, protein complexes belong to two categories: homomers, formed from multiple copies of the same protein subunit, and heteromers, which have at least two distinct subunits with different amino-acid sequences. While homomers and heteromers are both prevalent across evolution, most prokaryotic complexes are homomers, while most eukaryotic complexes are heteromers [5–7].

Protein complexes are crucial for a large number of biological functions, and different types of protein quaternary structures have been shown to facilitate different biological functions and allosteric regulation [8•,9–12]. A large number of other benefits have been proposed [4••,13]. For example, considering the possibility of acquiring mutations during transcription and translation, it is more efficient to synthesize a larger structure in modules of subunits. Importantly, it also allows fine spatial and temporal regulation, and reduces folding complexity in forming unique shapes such rings or filaments. It has also been shown that multiple identical domains of the same polypeptide chain are prone to aggregation [14] due to formation of domain-swapped structures during cotranslational folding [15•]. Therefore, translating these domains as separate polypeptides that later assemble into a large complex can be less risky. Finally, it is important to emphasize that, while clearly there are many advantages to protein complexes, protein oligomerization is not always functionally beneficial and the result of evolutionary selection, but may be explained by simple nonadaptive processes [6,16].

In recent years, we have learned a considerable amount about the processes by which proteins assemble into complexes. We know that proteins generally assemble via ordered pathways that tend to be evolutionarily conserved [17,18]. Moreover, these assembly pathways appear to be biologically important both in prokaryotes [19] and eukaryotes [20]. However, there are still unanswered questions about how the cell regulates protein complex assembly, and where assembly actually occurs within the cell. A logical place to begin addressing this is in the initial stages of protein synthesis and folding.

Cotranslational folding and assembly

The phenomenon of cotranslational folding has received considerable attention in recent years. Although the exact frequency at which cotranslational folding occurs in either prokaryotes or eukaryotes is unknown, there is a large body of computational [21–23] and experimental work [24,25••,26,27••] supporting and defining its likelihood. Significantly, these works emphasize the balance between the rate of translation, for example, as a function of charged-tRNA availability [28] or mRNA secondary structure [29–31], and the rate of protein folding. For reviews on the topic we recommend [32–34].

There are several reasons why proteins might acquire secondary structure during translation, sometimes even while still inside the ribosome exit tunnel [24,35,36]. For example, folding cotranslationally can modify the potential energy landscape to avoid nonproductive intermediates that would prevent the protein from reaching its native state [28]. However, cotranslational folding also reduces the propensity of deleterious non-specific interactions with the crowded cellular milieu or with other polypeptides on the same polyribosome. In other words, the protein primarily folds to protect itself from nonspecific interactions, but in doing so also allows assembly with native partners.

Given the prevalence of cotranslational folding, it is natural to imagine that assembly could also occur cotranslationally, especially given that folding and assembly are so intimately related [37]. This could potentially be beneficial for many of the same reasons as cotranslational folding; in particular, it could protect the protein from non-specific interactions, which is crucial due to the presence of the exposed interfaces making the unassembled subunits very sensitive to aggregation. This is particularly true for soluble homomers, which typically form larger hydrophobic interfaces than heteromers, and are thus more prone to misinteraction [38]. Although cotranslational assembly has received far less attention than cotranslational folding, it has been known of for a long time, with the first example we are aware of being homotetrameric β-galactosidase published in 1964 [39]. More recently, evidence is emerging that the phenomenon may be widespread [34,40••].

How does cotranslational assembly occur within the cell?

During cotranslational assembly, at least one of the protein subunits begins to assemble while it is still in the process of being translated, that is, the interaction involves a nascent chain. This can occur via either *cis* or *trans* mechanisms. The *cis* mechanism (Figure 1a) involves the assembly of polypeptides from the same mRNA; this can refer either to the case where an interaction occurs while both chains are still in the process of being translated, or when a nascent chain binds to a fully translated protein released by the same mRNA. In contrast, the *trans* mechanism (Figure 1b) involves the assembly of a polypeptide from one mRNA with the product of another, and can apply to either heteromeric or homomeric assembly.

The rate at which cotranslational assembly will occur is a function of the affinity of the subunits for one another, and their effective concentration. However, concentration in this case is not purely determined by the number of proteins in solution, but also by the density of nascent polypeptides on the polyribosome. An important parameter influencing this is the length of time a nascent polypeptide spends attached to the mRNA, which in turn depends on numerous factors, including mRNA secondary structure [30], the availability of charged-tRNAs, the overall length of the mRNA, and elements such as anti-Shine-Dalgarno sequences in mRNA [41]. Thus, concentration is a function of multiple variables, but for simplicity can be summarized as the total number of nascent polypeptides within the polyribosome's sphere of influence at a particular point in time.

At this point, we would like to propose an additional role to the secondary structure of mRNA. As mentioned above, the secondary structure of mRNA affects translation rate, thus regulating nascent chain folding into its correct fold. However, it is likely that many mRNAs form more complex structures than that of the two-dimensional structure, and thus the polyribosome and consequently the ribosome tunnels will be orientated in a particular way. These trajectories will influence both the probability of clashing between nascent chains, which will affect the stability of monomers, and the probability of cotranslational complex assembly. It is therefore important to understand the native three-dimensional of the polyribosome, continuing recent efforts [42••,43•].

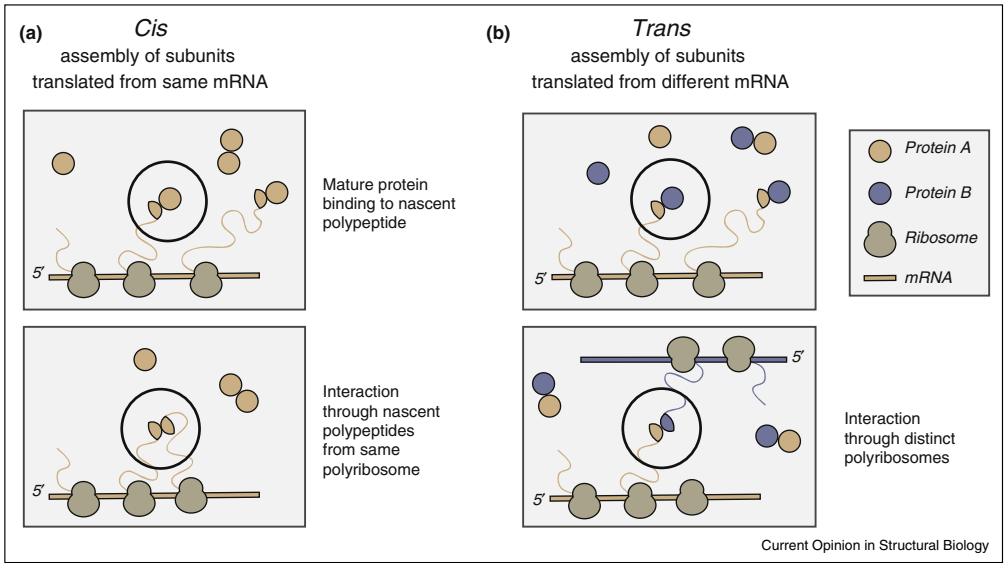
The cell broadly regulates both *cis* and *trans* mechanisms. For *cis*, the number of ribosomes, which is a function of 'initiation rate' (how many), 'elongation' (how long), and 'termination', will determine its frequency of occurrence. For the *trans* mechanism, concentration can be increased by active transport of the same-gene mRNAs transcripts to a specific location in the cell, a mechanism which has been observed in both eukaryotes [44] and prokaryotes [45,46]. It is worth mentioning that this factor is rarely discussed in the literature, and should be taken into account while discussing mRNA localization of protein complexes.

Cotranslational assembly of operon-encoded complexes

At this juncture, it is important to highlight the stark differences between eukaryotic and prokaryotic assembly of protein complexes, specifically for heteromers. In eukaryotes, cotranslational assembly of heteromers must occur in *trans*, either through co-localization of mRNAs encoding interacting proteins, or through localization of fully folded proteins to active polysomes (Figure 1b). In contrast, prokaryotes often encode protein complex subunits in operons, whereby distinct protein subunits can be translated from the same polycistronic mRNA molecule [47,48]. Thus, for operon-encoded complexes, cotranslational assembly of heteromers can occur in *cis* in much the same way as it does for homomers (Figure 2).

To this end, there are multiple strands of evidence pointing to the important role operons play in facilitating

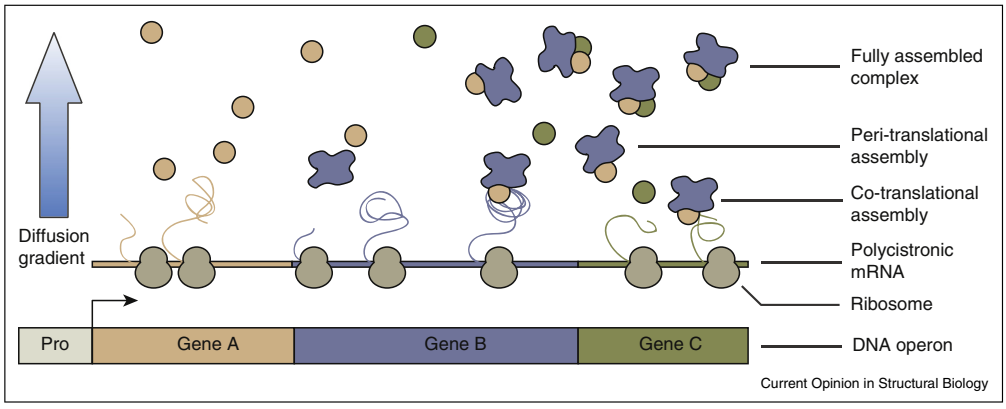
Figure 1



Cotranslational assembly can occur via **(a)** *cis* or **(b)** *trans* mechanisms, in which the interacting subunits are translated from either the same or different mRNA molecules. Moreover, either one or both subunits may still be in the process of being translated when the interaction occurs. *Cis* assembly exclusively involves homomers, whereas *trans* assembly can involve either two subunits encoded by the same gene (homomers) or different genes (heteromers).

complex assembly. In recent work using a modified luciferase system, Shieh *et al.* [25••] demonstrated that encoding the genes for LuxA and LuxB within a single operon leads to markedly improved assembly efficiency compared to encoding them in different operons. They were further able to show directly that interactions between LuxA and LuxB were being formed cotranslationally. This was achieved by co-purifying YFP-tagged subunits with ribosomes that were actively translating untagged partner proteins.

Figure 2



Prokaryotic heteromers are often encoded in operons, whereby multiple genes are transcribed onto a single polycistronic mRNA. The order of genes in operons is highly non-random, and has been selected for such that adjacent genes on the operon are more likely to physically interact as proteins. Similarly, the order in which the protein complex assembles typically reflects the order in which the genes are encoded. This implies that assembly occurs cotranslationally, facilitated by the close proximity of interacting subunits. Further support comes from the fact that the correlation between gene order and assembly order is stronger for weakly expressed complexes, where it is essential that assembly takes place rapidly, before subunits diffuse away from the site of translation.

A complementary approach to the experimental work just described used computational analysis of structural and genomic data to demonstrate a strong correspondence between operon gene order and the assembly order of protein complexes, that is, proteins that are translated first tend be those that assemble first [19]. Moreover, adjacent genes in operons are far more likely to encode physically interacting proteins that form large interfaces than those separated by intervening DNA. For our purposes, the important implication arising from this is that these subunits must be assembling cotranslationally or very shortly after translation (i.e. peri-translationally). If not, then any selection for gene order would be rendered effectively neutral due to the diffusion away from the site of translation occurring prior to assembly. These studies, along with reports of increased yield in protein complexes from using native operon order when designing expression vectors [49], make it clear that cotranslational assembly of heteromers must be widespread in prokaryotes.

Factors influencing cotranslational assembly and its influence on protein complex evolution

Clearly there are advantages to cotranslational assembly, such as misinteraction avoidance and speed of assembly, but are there any drawbacks that might limit its occurrence in nature? One such drawback was first demonstrated by Jaenicke and colleagues [50–53], who showed that *in vitro* refolding of homomeric proteins after denaturation is more challenging than it is for monomeric proteins, presumably due to premature assembly [54]. Here we highlight a few additional scenarios in which cotranslational assembly may have deleterious effects.

First, assembly may slow or even pause ribosomes from their rapid unidirectional sliding along mRNA [55]. Second, assembly constrains the freedom of the nascent chain to freely rotate in all three rotational axes in the quest for the native fold. Limiting the polypeptide's rotational freedom may in fact direct the protein to the correct fold, that is, by limiting undesirable folds, but that may not be the case for all proteins. For example, knotted proteins, unique topological structures that form via the thread of one terminus through a loop of an intermediate conformer [56], are likely to avoid cotranslational assembly. Last, *cis* cotranslational assembly may force high proximity between two (or more) unfolded nascent chains; in other words, upon assembly a triangle-like conformation is adopted by the chains, with the tip of the triangle being the assembly point connecting two partially unfolding nascent chains. This premature assembly scenario could also explain the *in vitro* work of Jaenicke and colleagues.

Following this line of work, we hypothesized that cotranslational assembly is likely to be limited by different constraints because of the unique situation cotranslational assembly forces upon the nascent chains; that is, the linking of these molecules in the midst dynamic elongation and folding processes. Therefore, we performed a combined computational and experimental analysis to investigate this phenomenon [57]. First, we observed highly significant trend for interface-forming residues in homomers to be located towards C termini across thousands of protein structures and diverse kingdoms of life. This was in contrast to heteromers, where no such tendency was observed. We suspect this trend is the result of cotranslational assembly being evolutionarily selected against under certain circumstances: localization of interfaces towards C-termini will reduce the chance of cotranslational assembly since interface-forming residues will be translated last. To address this further, we expressed all homomers of *Escherichia coli* with known structures and assessed them for their *in vivo* aggregation propensities. Interestingly, the results showed that homomers with N-terminal interfaces are more likely to show an early and severe aggregation, supporting the idea that cotranslational interactions between homomeric subunits can lead to protein misfolding and misassembly.

We also investigated the factors that allow successful cotranslational assembly by engineering a library of constructs comprising three components organized in different orders: first, oligomerization domain that folds cotranslationally, second, a linker, and third, reporter genes. The position of the oligomerization domain was critical for the stability of the protein: positioning it at the N terminus results in misassembly, which correlates with the propensity for assembly to occur cotranslationally, in comparison to the well-folded C-terminal variant. However, successful assembly can still occur via the N terminus if a linker extends the distance between the oligomerization domain and the reporter, which suggests that the increase of the linker could either decrease local concentration and thus the propensity to assemble. Alternatively, if cotranslational assembly did occur, the local concentration of the partially unfolded nascent chains is reduced, thus lowering the propensity for misassembly. Finally, increasing the reporter's folding rate also allows successful cotranslational assembly via the N terminus, suggesting that enhanced folding of a domain adjacent to the assembly site increases the probability for protein stability. This finding may also align with the notion of extreme proximity of nascent chains upon assembly, whereby acquiring secondary structures fast enough protects the polypeptide from non-specific interactions.

This is the first work to our knowledge to describe the parameters by which cotranslational assembly works. However, it mainly focused on mechanisms encoded in the protein primary sequence, such as the location of residues participating in assembly or protein folding rate. Clearly, other factors such as chaperones may participate in ensuring correct assembly both for homomers and heteromers.

The role of cotranslational chaperones in regulating assembly

Chaperones play an essential role in avoiding misfolding or aggregation, thus promoting the formation of native tertiary and quaternary protein structure. The mechanistic details of how they act vary dramatically, and chaperones as a whole encompass a wide variety of unrelated protein families. There are several chaperones that directly assist the assembly of protein complexes. For example: the PAC family, which form intra-family heterodimers that assist with the assembly of heptameric alpha-subunit rings in the proteasome [58].

However, chaperones more often facilitate assembly indirectly, by ensuring that unfolded proteins reach their native-state safely, thus allowing correct assembly later [59]. Proteins are most vulnerable to formation of non-specific interactions during the process of translation, and thus it is unsurprising that many of these chaperones themselves act cotranslationally. For example, Hsp70 family members, together with Hsp40 co-chaperones, can interact cotranslationally with nascent polypeptide chains, protecting them against premature misfolding and aggregation [60]. Similarly, TRiC and the prokaryotic Trigger factor act downstream, facilitating folding and oligomeric assembly [61].

The action of chaperones is particularly important for eukaryotic proteins, which are typically longer than those from prokaryotes, often comprise multiple domains, and have a higher incidence of intrinsically disordered and flexible regions [7,62,63], which is in stark comparison to prokaryote proteins that shift the folding process towards a posttranslational route ([60] and references therein). The implication for our discussion is that we should expect to find more examples of chaperone involvement, directly or indirectly, in cotranslational assembly.

A final intriguing case is that of cotranslational interaction between human mitochondrially encoded COX1 and C12ORF62 [64]. COX1 is the first subunit of cytochrome c oxidase (complex IV of the respiratory chain complexes). During translation by the mitochondrial ribosome, it associates cotranslationally with two membrane-embedded assembly factors: first C12ORF62 and then MITRAC12. This enables interaction with the nuclear-encoded COX4, which is the second complex IV subunit to bind. Crucially, COX4 is the trigger for the release of COX1 by the ribosome. In COX4-depleted cells, the nascent COX1-C12ORF62 intermediate is held in an 'assembly-primed' state and simply accumulates in the mitochondrial inner membrane. As a result, those mitochondrial ribosomes translating COX1 are prevented from creating further copies of COX1. The mechanistic details of this process are not yet fully understood, but it has a fascinating implication, namely that mitochondrial translation activity can react to changes in the production of nuclear-encoded proteins. When cytoplasmic production of complex IV subunits slows, so too does mitochondrial production, despite the fact that the subunits in question are encoded on different genomes.

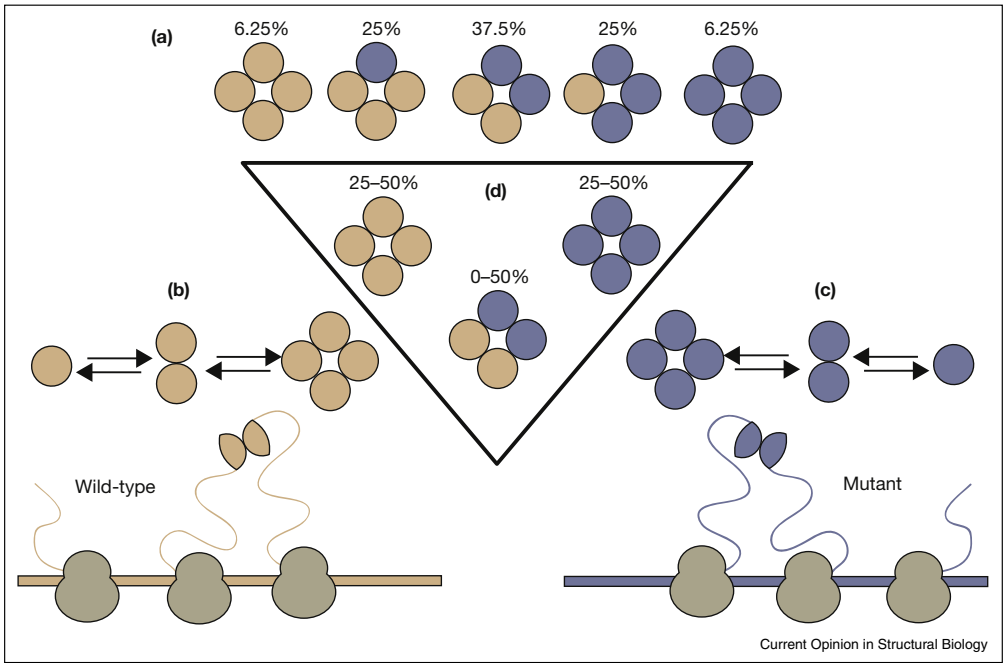
Implications of cotranslational assembly for the inheritance of genetic disease

The phenomenon of cotranslational assembly is not just important for understanding protein complex regulation and evolution: it also has potentially very important implications for genetic disorders associated with a dominant-negative (DN) mode of action. Essentially, a DN effect occurs when expression of a mutant allele can disrupt the activity of the wild type allele [65], thus resulting in a dominant mode of inheritance. DN effects have often been seen for genes that encode proteins that assemble into homomers [66]. The reason for this is simple: if the presence of a single mutant subunit within a complex is enough to 'poison' of the complex, the result will be a far greater reduction in activity than the 50% expected for a simple heterozygous loss-of-function mutation. In fact, DN mutations tend to be significantly less destabilizing towards protein structure than other pathogenic mutations because the mechanism requires that complex is still able to assemble [67].

To illustrate this, we can consider the case of a homotetramer encoded by a heterozygous allele. If both subunits are expressed at equal levels and associate randomly, then only 1/16 (6.25%) of the assembled complexes will be fully wild type homomers (Figure 3a). In contrast, if assembly occurs in *cis*, that is, cotranslationally or peri-translationally, the stoichiometry of the assembled complexes will be different. If all assembly occurs in *cis*, the homomeric products will be homogeneous, with half of the assembled complexes being fully wild type (Figure 3b) and half being fully mutant (Figure 3c). Finally, if not all of the second assembly step (dimerization of dimers) occurs peri-translationally, or there is equilibrium exchange between tetrameric and dimeric states, then the proportion of full wild-type complex will be smaller, but still greater than in the case of totally random assembly (Figure 3d). Therefore, the phenomenon of cotranslational assembly should reduce the likelihood that a DN mechanism of pathogenesis will be observed, since the proportion of homogeneous wild-type complex will be greater.

Importantly, the dissociation constant of the complex also plays a role in the final 'mixing' with other alleles once diffused away from the polyribosome. For example, the p53 homotetramer, was found to dimerize cotranslationally [68], which ensures that the complex is unlikely to form mixed primary dimers in the protein's short lifetime, promoting a better mixing strategy in the case of DN

Figure 3



Cotranslational assembly of a homomer encoded by a heterozygous allele affects the stoichiometry of assembled complexes and can influence the dominant-negative mechanism of molecular inheritance. If wild-type and mutant subunits associate randomly, the distribution of stoichiometries in **(a)** will be seen, and only 1/16 (6.25%) complexes will be fully wild type. If assembly is completely co- or peri-translational, then the assembles complexes will contain either **(b)** all wild-type or **(c)** all mutant subunits. Finally, if the second assembly step (dimerization of dimers) is not obligately cotranslational, or there is a conformational equilibrium between tetramers and dimers, then the stoichiometries of assembled complexes can be within the ranges shown in **(d)**.

mutations. Some of p53 mutations indeed behave in a DN fashion: mostly structural mutations that can enhance aggregation of wild type that co-exists in the same tetrameric complex. However, the deleterious effect of many mutations can in fact be diluted by the wild type [69•], which may explain why some tumours discard the wild-type allele [70].

Concluding remark

Assembly of protein complexes often occurs very close to the site of translation. This is due to effects of cellular crowding, which limits diffusion, and significantly reduces the probability of lowly expressed subunits finding their binding partners outside of the high local concentrations surrounding the ribosome. Moreover, such an assembly limits the time of uncovered hydrophobic interfaces that makes the unassembled subunits very sensitive to aggregation. Peripheral assembly will also determine the composition of the complex, considering the presence of disease forming alleles. Nevertheless, cotranslational assembly can also carry a heavy cost, namely through the formation of aggregates, whether non-specific or amyloid. As such, the cell must strike a balance between rapid assembly near the ribosome and avoidance of aggregation that ensures the stability of the polypeptide's tertiary and quaternary structure.

To support these ideas, and to further understand the role of cotranslational assembly in normal biological function, as well as its potential implications mitigating the DN effect in inherited and *de novo* genetic disorders, there is a need for new tools and much more experimental characterization cotranslational processes. For example, NMR [71,72], cryoelectron microscopy [42••] and proteomics [40••] have shown great promise, and are likely to continue to do so in coming years.

Conflict of interest

Nothing declared.

Acknowledgement

J.A.M. is supported by a Medical Research Council Career Development Award (MR/M02122X/1).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Marsh JA, Teichmann SA: **Structure, dynamics, assembly, and evolution of protein complexes**. *Annu Rev Biochem* 2015, **84**:551-575.

2. • Hein MY, Hubner NC, Poser I, Cox J, Nagaraj N, Toyoda Y, Gak IA, Weisswange I, Mansfeld J, Buchholz F *et al.*: **A human interactome in three quantitative dimensions organized by stoichiometries and abundances**. *Cell* 2015, **163**:712-723.

Using a large library of GFP-tagged proteins under near-endogenous control, the authors were able to describe the human interactome including specificities, stoichiometries, and abundances.

3. Ahnert SE, Marsh JA, Hernández H, Robinson CV, Teichmann SA: **Principles of assembly reveal a periodic table of protein complexes**. *Science* 2015, **350**:aaa2245.

4. •• Goodsell DS, Olson AJ: **Structural symmetry and protein function**. *Annu Rev Biophys Biomol Struct* 2000, **29**:105-153.

A fundamental and thorough review in the field of protein complexes describing both soluble and membrane-bound oligomeric proteins, including various evolutionary factors such as symmetry and size.

5. Marsh JA, Rees HA, Ahnert SE, Teichmann SA: **Structural and evolutionary versatility in protein complexes with uneven stoichiometry**. *Nat Commun* 2015, **6**:6394.

6. Lynch M: **The evolution of multimeric protein assemblages**. *Mol Biol Evol* 2012, **29**:1353-1366.

7. Marsh JA, Teichmann SA: **Protein flexibility facilitates quaternary structure assembly and evolution**. *PLoS Biol* 2014, **12**:e1001870.

8. • Pieters BJGE, van Eldijk MB, Nolte RJM, Mecinović J: **Natural supramolecular protein assemblies**. *Chem Soc Rev* 2016, **45**:24-39.

A recent review that well describes the structure–function relationship of different types protein complexes.

9. Forrest LR: **Structural symmetry in membrane proteins**. *Annu Rev Biophys* 2015, **44**:311-337.

10. Bergendahl T, Marsh JA: **Functional determinants of protein assembly into homomeric complexes**. *bioRxiv* 2016 http://dx.doi.org/10.1101/081745.

11. Changeux J-P: **Allostery and the Monod–Wyman–Changeux model after 50 years**. *Annu Rev Biophys* 2012, **41**:103-133.

12. Marianayagam NJ, Sunde M, Matthews JM: **The power of two: protein dimerization in biology**. *Trends Biochem Sci* 2004, **29**:618-625.

13. Ali MH, Imperiali B: **Protein oligomerization: how and why**. *Bioorg Med Chem* 2005, **13**:5013-5020.

14. Wright CF, Teichmann SA, Clarke J, Dobson CM: **The importance of sequence diversity in the aggregation and evolution of proteins**. *Nature* 2005, **438**:878-881.

15. • Borgia MB, Borgia A, Best RB, Steward A, Nettels D, Wunderlich B, Schuler B, Clarke J: **Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins**. *Nature* 2011, **474**:662-665.

Using single-molecule FRET, the authors quantify interdomain misfolding as a function of the domains' sequence identity via the formation of domain-swapped mechanism.

16. Lynch M: **Evolutionary diversification of the multimeric states of proteins**. *Proc Natl Acad Sci U S A* 2013, **110**:E2821-E2828.

17. Levy ED, Boeri Erba E, Robinson CV, Teichmann SA: **Assembly reflects evolution of protein complexes**. *Nature* 2008, **453**:1262-1265.

18. Marsh JA, Hernández H, Hall Z, Ahnert SE, Perica T, Robinson CV, Teichmann SA: **Protein complexes are under evolutionary selection to assemble via ordered pathways**. *Cell* 2013, **153**:461-470.

19. Wells JN, Bergendahl LT, Marsh JA: **Operon gene order is optimized for ordered protein complex assembly**. *Cell Rep* 2016, **14**:679-685.

20. McShane E, Sin C, Zauber H, Wells JN, Donnelly N, Wang X, Hou J, Chen W, Storchova Z, Marsh JA *et al.*: **Kinetic analysis of protein stability reveals age-dependent degradation**. *Cell* 2016, **167**:803-815.

21. Sharma AK, Bukau B, O'Brien EP: **Physical origins of codon positions that strongly influence cotranslational folding: a framework for controlling nascent-protein folding**. *J Am Chem Soc* 2016, **138**:1180-1195.

22. O'Brien EP, Vendruscolo M, Dobson CM: **Kinetic modelling indicates that fast-translating codons can coordinate cotranslational protein folding by avoiding misfolded intermediates**. *Nat Commun* 2014, **5**:2988.

23. O'Brien EP, Ciryam P, Vendruscolo M, Dobson CM: **Understanding the influence of codon translation rates on cotranslational protein folding**. *Acc Chem Res* 2014, **47**:1536-1544.

24. Nilsson OB, Hedman R, Marino J, Wickles S, Bischoff L, Johansson M, Müller-Lucks A, Trovato F, Puglisi JD, O'Brien EP *et al.*: **Cotranslational protein folding inside the ribosome exit tunnel**. *Cell Rep* 2015, **12**:1533-1540.

25. •• Shieh Y-W, Minguez P, Bork P, Auburger JJ, Guilbride DL, Kramer G, Bukau B: **Operon structure and cotranslational subunit association direct protein assembly in bacteria**. *Science* 2015, **350**:678-680.

Elegant demonstration of cotranslational assembly in operon-encoded protein complexes, also showing that encoding subunits within operons leads to a marked increase in complex assembly efficiency.

26. Buhr F, Jha S, Thommen M, Mittelstaet J, Kutz F, Schwalbe H, Rodnina MV, Komar AA: **Synonymous codons direct cotranslational folding toward different protein conformations**. *Mol Cell* 2016, **61**:341-351.

27. •• Sander IM, Chaney JL, Clark PL: **Expanding Anfinsen's principle: contributions of synonymous codon selection to rational protein design**. *J Am Chem Soc* 2014, **136**:858-861.

Elegant work showing the effects of cotranslational folding on the structure of the encoded protein *in vivo*.

28. Zhang G, Ignatova Z: **Folding at the birth of the nascent chain: coordinating translation with co-translational folding**. *Curr Opin Struct Biol* 2011, **21**:25-31.

29. Faure G, Ogurtsov AY, Shabalina SA, Koonin EV: **Role of mRNA structure in the control of protein folding**. *Nucleic Acids Res* 2016 http://dx.doi.org/10.1093/nar/gkw671.

30. Endoh T, Sugimoto N: **Mechanical insights into ribosomal progression overcoming RNA G-quadruplex from periodical translation suppression in cells**. *Sci Rep* 2016, **6**:22719.

31. Endoh T, Kawasaki Y, Sugimoto N: **Synchronized translation for detection of temporal stalling of ribosome during single-turnover translation**. *Anal Chem* 2012, **84**:857-861.

32. Nissley DA, O'Brien EP: **Timing is everything: unifying codon translation rates and nascent proteome behavior**. *J Am Chem Soc* 2014, **136**:17892-17898.

33. Jacobson GN, Clark PL: **Quality over quantity: optimizing co-translational protein folding with non-'optimal' synonymous codons**. *Curr Opin Struct Biol* 2016, **38**:102-110.

34. Wells JN, Bergendahl LT, Marsh JA: **Co-translational assembly of protein complexes**. *Biochem Soc Trans* 2015, **43**:1221-1226.

35. Kosolapov A, Deutsch C: **Tertiary interactions within the ribosomal exit tunnel**. *Nat Struct Mol Biol* 2009, **16**:405-411.

36. Bhushan S, Gartmann M, Halic M, Armache J-P, Jarasch A, Mielke T, Berninghausen O, Wilson DN, Beckmann R: **alpha-Helical nascent polypeptide chains visualized within distinct regions of the ribosomal exit tunnel**. *Nat Struct Mol Biol* 2010, **17**:313-317.

37. Dyson HJ, Wright PE: **Coupling of folding and binding for unstructured proteins**. *Curr Opin Struct Biol* 2002, **12**:54-60.

38. Levy ED, Teichmann S: **Structural, evolutionary, and assembly principles of protein oligomerization**. *Prog Mol Biol Transl Sci* 2013, **117**:25-51.

39. Kiho Y, Rich A: **Induced enzyme formed on bacterial polyribosomes**. *Proc Natl Acad Sci U S A* 1964, **51**:111-118.

40. •• Duncan CDS, Mata J: **Widespread cotranslational formation of protein complexes**. *PLoS Genet* 2011, **7**:e1002398.
First work to provide evidence that cotranslational assembly of complexes is common in eukaryotic cells, thus generalising numerous individual examples from earlier literature

41. Li G-W, Oh E, Weissman JS: **The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria**. *Nature* 2012, **484**:538-541.

42. •• Brandt F, Etchells SA, Ortiz JO, Elcock AH, Hartl FU, Baumeister W: **The native 3D organization of bacterial polysomes**. *Cell* 2009, **136**:261-271.
Using cryoelectron tomography, the authors show polysomal arrangements for preferred orientation, potentially to avoid clashes between the nascent chains and allowing access of tRNA.

43. • Mrazek J, Toso D, Ryazantsev S, Zhang X, Zhou ZH, Fernandez BC, Kickhoefer VA, Rome LH: **Polyribosomes are molecular 3D nanoprinters that orchestrate the assembly of vault particles**. *ACS Nano* 2014, **8**:11552-11559.
Using electron microscopy, this exciting work shows the cotranslational assembly of the megadalton-size vault complex.

44. Martin KC, Ephrussi A: **mRNA localization: gene expression in the spatial dimension**. *Cell* 2009, **136**:719-730.

45. Montero Llopis P, Jackson AF, Sliusarenko O, Surovtsev I, Heinritz J, Emonet T, Jacobs-Wagner C: **Spatial organization of the flow of genetic information in bacteria**. *Nature* 2010, **466**:77-81.

46. Shapiro L, McAdams HH, Losick R: **Why and how bacteria localize proteins**. *Science* 2009, **326**:1225-1228.

47. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact**. *Trends Biochem Sci* 1998, **23**:324-328.

48. Mushegian AR, Koonin EV: **Gene order is not conserved in bacterial evolution**. *Trends Genet* 1996, **12**:289-290.

49. Poulsen C, Holton S, Geerlof A, Wilmanns M, Song Y-H: **Stoichiometric protein complex formation and over-expression using the prokaryotic native operon structure**. *FEBS Lett* 2010, **584**:669-674.

50. Jaenicke R, Seckler R: **Protein misassembly in vitro**. *Adv Protein Chem* 1997, **50**:1-59.

51. Jaenicke R: **Protein folding: local structures, domains, subunits, and assemblies**. *Biochemistry* 1991, **30**:3147-3161.

52. Seckler R, Fuchs A, King J, Jaenicke R: **Reconstitution of the thermostable trimeric phage P22 tailspike protein from denatured chains in vitro**. *J Biol Chem* 1989, **264**:11750-11753.

53. Rudolph R, Zettlmeissl G, Jaenicke R: **Reconstitution of lactic dehydrogenase. Noncovalent aggregation vs. reactivation. 2. Reactivation of irreversibly denatured aggregates**. *Biochemistry* 1979, **18**:5572-5575.

54. Jaenicke R, Lilie H: **Folding and association of oligomeric and multimeric proteins**. *Adv Protein Chem* 2000, **53**:329-401.

55. Proshkin S, Rahmouni AR, Mironov A, Nudler E: **Cooperation between translating ribosomes and RNA polymerase in transcription elongation**. *Science* 2010, **328**:504-508.

56. Lim NCH, Jackson SE: **Mechanistic insights into the folding of knotted proteins in vitro and in vivo**. *J Mol Biol* 2015, **427**:248-258.

57. Natan E, Endoh T, Haim-Vilmovsky L, Chalancon G, Flock T, Hopper JTS, Kintses B, Daruka L, Fekete G, Pal C *et al.*: **Cotranslational assembly imposes evolutionary constraints on homomeric proteins**. *bioRxiv* 2016 http://dx.doi.org/10.1101/074963.

58. Ramos PC, Dohmen RJ: **PACemakers of proteasome core particle assembly**. *Structure* 2008, **16**:1296-1304.

59. Makhnevych T, Houry WA: **The role of Hsp90 in protein complex assembly**. *Biochim Biophys Acta* 2012, **1823**:674-682.

60. Willmund F, del Alamo M, Pechmann S, Chen T, Albanèse V, Dammer EB, Peng J, Frydman J: **The cotranslational function of ribosome-associated Hsp70 in eukaryotic protein homeostasis**. *Cell* 2013, **152**:196-209.

61. Hoffmann A, Bukau B, Kramer G: **Structure and function of the molecular chaperone Trigger Factor**. *Biochim Biophys Acta* 2010, **1803**:650-661.

62. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT: **Prediction and functional analysis of native disorder in proteins from the three kingdoms of life**. *J Mol Biol* 2004, **337**:635-645.

63. Schad E, Tompa P, Hegyi H: **The relationship between proteome size, structural disorder and organism complexity**. *Genome Biol* 2011, **12**:R120.

64. Richter-Dennerlein R, Oeljeklaus S, Lorenzi I, Ronsör C, Bareth B, Schendzielorz AB, Wang C, Warscheid B, Rehling P, Dennerlein S: **Mitochondrial protein synthesis adapts to influx of nuclear-encoded protein**. *Cell* 2016, **167** 471–83.e10.

65. Herskowitz I: **Functional inactivation of genes by dominant negative mutations**. *Nature* 1987, **329**:219-222.

66. Veitia RA: **Exploring the molecular etiology of dominant-negative mutations**. *Plant Cell* 2007, **19**:3843-3851.

67. McEntagart M, Williamson KA, Rainger JK, Wheeler A, Seawright A, De Baere E, Verdin H, Bergendahl LT, Quigley A, Rainger J *et al.*: **A restricted repertoire of de novo mutations in ITPR1 cause Gillespie syndrome with evidence for dominant-negative effect**. *Am J Hum Genet* 2016, **98**:981-992.

68. Nicholls CD, McLure KG, Shields MA, Lee PWK: **Biogenesis of p53 involves cotranslational dimerization of monomers and posttranslational dimerization of dimers. Implications on the dominant negative effect**. *J Biol Chem* 2002, **277**:12937-12945.

69. • Perica T, Marsh JA, Sousa FL, Natan E, Colwell LJ, Ahnert SE, Teichmann SA: **The emergence of protein complexes: quaternary structure, dynamics and allostery. Colworth Medal Lecture**. *Biochem Soc Trans* 2012, **40**:475-491.
A review that described the evolutionary aspects of protein complexes. Importantly, the authors discuss the role of oligomerization in the dominant-negative effects, suggesting that oligomerization may evolve to dilute mutations' effects.

70. Olive KP, Tuveson DA, Ruhe ZC, Yin B, Willis NA, Bronson RT, Crowley D, Jacks T: **Mutant p53 gain of function in two mouse models of Li-Fraumeni syndrome**. *Cell* 2004, **119**:847-860.

71. Deckert A, Waudby CA, Wlodarski T, Wentink AS, Wang X, Kirkpatrick JP, Paton JFS, Camilloni C, Kukic P, Dobson CM *et al.*: **Structural characterization of the interaction of α-synuclein nascent chains with the ribosomal surface and trigger factor**. *Proc Natl Acad Sci U S A* 2016, **113**:5012-5017.

72. Waudby CA, Launay H, Cabrita LD, Christodoulou J: **Protein folding on the ribosome studied using NMR spectroscopy**. *Prog Nucl Magn Reson Spectrosc* 2013, **74**:57-75.

BIBLIOGRAPHY

1. Wells, J. N., Bergendahl, L. T. & Marsh, J. A. Operon Gene Order Is Optimized for Ordered Protein Complex Assembly. *Cell Reports* **14,** 679–685. ISSN: 22111247 (Feb. 2016).
2. McShane, E. *et al.* Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. *Cell* **167,** 803–815. ISSN: 00928674 (Oct. 2016).
3. Wells, J. N., Gligoris, T. G., Nasmyth, K. A. & Marsh, J. A. Evolution of condensin and cohesin complexes driven by replacement of Kite by Hawk proteins. *Current Biology* **27,** R17–R18. ISSN: 09609822 (Jan. 2017).
4. Larson, S. M., England, J. L., Desjarlais, J. R. & Pande, V. S. Thoroughly sampling sequence space: Large-scale protein design of structural ensembles. *Protein Science* **11,** 2804–2813. ISSN: 09618368 (Apr. 2002).
5. Povolotskaya, I. S. & Kondrashov, F. A. Sequence space and the ongoing expansion of the protein universe. *Nature* **465,** 922–926. ISSN: 0028-0836 (2010).
6. Lynch, M. Evolutionary diversification of the multimeric states of proteins. *Proceedings of the National Academy of Sciences of the United States of America* **110,** E2821–8. ISSN: 1091-6490 (July 2013).
7. Dayhoff, J. E., Shoemaker, B. A., Bryant, S. H. & Panchenko, A. R. Evolution of Protein Binding Modes in Homooligomers. *Journal of Molecular Biology* **395,** 860–870. ISSN: 00222836 (Jan. 2010).
8. Alexander, E., Bergendahl, L. T., Vincent, S., Marsh, J. A. & Warnecke, T. The genetic basis and evolution of red blood cell sickling in deer. *BioArxiv* (2017).
9. Garcia-Seisdedos, H., Empereur-Mot, C., Elad, N. & Levy, E. D. Proteins evolve on the edge of supramolecular self-assembly. *Nature.* ISSN: 0028-0836. doi:10.1038/nature23320. http://www.nature.com/doifinder/10.1038/nature23320 (Aug. 2017).
10. Gould, S. J. & Lewontin, R. C. The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proceedings of the Royal Society of London. Series B, Biological Sciences* **205,** 581–598 (1979).
11. Koonin, E. V. Splendor and misery of adaptation , or the importance of neutral null for understanding evolution. *BMC Biology,* 1–8. ISSN: 1741-7007 (2016).
12. Gingold, H. & Pilpel, Y. Determinants of translation efficiency and accuracy. *Molecular Systems Biology* **7,** 481. ISSN: 1744-4292 (2011).

13. Monod, J., Changeux, J.-P. & Jacob, F. Allosteric proteins and cellular control systems. *Journal of Molecular Biology* **6,** 306–329. ISSN: 00222836 (1963).

14. Perutz, M. F. Structure and mechanism of haemoglobin. *British Medical Bulletin* **32,** 195–208. ISSN: 0007-1420 (1976).

15. Kolatkar, P. R., Meador, W. E., Stanfield, R. L. & Hackert, M. L. Novel subunit structure observed for noncooperative hemoglobin from Urechis caupo. *Journal of Biological Chemistry* **263,** 3462–3465. ISSN: 00219258 (1988).

16. Royer, W. E., Zhu, H., Gorr, T. A., Flores, J. F. & Knapp, J. E. Allosteric Hemoglobin Assembly: Diversity and Similarity. *Journal of Biological Chemistry* **280,** 27477–27480. ISSN: 0021-9258 (July 2005).

17. Bellelli, A. & Brunori, M. Hemoglobin allostery: Variations on the theme. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1807,** 1262–1272. ISSN: 00052728 (Oct. 2011).

18. Stanley, W. M. Isolation of a crystalline protein posessing the properties of tobacco-mosaic virus. *Science* **81,** 644–645 (1935).

19. Schrödinger, E. *What is life? The physical aspect of the living cell* 194. ISBN: 0-521-42708-8 (Cambridge University Press, 1947).

20. Dronamraju, K. R. Erwin Schrödinger and the origins of molecular biology. *Genetics* **153,** 1071–6. ISSN: 0016-6731 (Nov. 1999).

21. Fraenkel-Conrat, H. & C.Williams, R. Reconstitution of Active Tobacco Mosaic Virus From Its Inactive Protein and Nucleic Acid Components. *Proceedings of the National Academy of Sciences* **41,** 690–698. ISSN: 0027-8424 (1955).

22. Kendrew, J. C. *et al.* A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **181,** 662–666 (1958).

23. Perutz, M. F. *et al.* Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. *Nature* **185,** 416–422 (1960).

24. Schluenzen, F. *et al.* Structure of Functionally Activated Small Ribosomal Subunit at 3.3 Å Resolution. *Cell* **102,** 615–623. ISSN: 00928674 (Sept. 2000).

25. Ramakrishnan, V. *et al.* Structure of the 30S ribosomal subunit. *Nature* **407,** 327–339. ISSN: 00280836 (Sept. 2000).

26. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289,** 905–20. ISSN: 0036-8075 (Aug. 2000).

27. Fields, S. & Song, O.-k. A novel genetic system to detect protein-protein interactions. *Nature* **340,** 245–246. ISSN: 0028-0836 (July 1989).

28. Rajagopala, S. V. *et al.* The binary protein-protein interaction landscape of Escherichia coli. *Nat Biotech* **32,** 285–290. ISSN: 1087-0156 (Mar. 2014).

29. Karas, M., Bachmann, D. & Hillenkamp, F. Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. *Analytical Chemistry* **57,** 2935–2939. ISSN: 0003-2700 (Dec. 1985).

30. Tanaka, K. *et al.* Protein and polymer analyses up tom/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* **2,** 151–153. ISSN: 0951-4198 (Aug. 1988).

31. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71. ISSN: 0036-8075, 1095-9203 (1989).

32. Margaret O. Dayhoff. *Atlas of Protein Sequence and Structure* Vol. 1 (1965).

33. Bixon, M. & Lifson, S. Potential functions and conformations in cycloalkanes. *Tetrahedron* **23,** 769–784. ISSN: 00404020 (1967).

34. Levitt, M. The birth of computational structural biology. *Nature structural biology* **8,** 392–393. ISSN: 1072-8368 (2001).

35. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry* **30,** 1545–1614. ISSN: 01928651 (2009).

36. Salomon-Ferrer, R., Case, D. A. & Walker, R. C. An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **3,** 198–210. ISSN: 17590876 (Mar. 2013).

37. Case, D. *et al. AMBER 2017* 2017.

38. Rost, B., Fariselli, P. & Casadio, R. Topology prediction for helical transmembrane proteins at 86% accuracy-Topology prediction at 86% accuracy. *Protein Science* **5,** 1704–1718. ISSN: 09618368 (Aug. 1996).

39. Xu, Y., Xu, D. & Uberbacher, E. C. An efficient computational method for globally optimal threading. *Journal of computational biology : a journal of computational molecular cell biology* **5,** 597–614. ISSN: 1066-5277 (Jan. 1998).

40. Link, A. J., Fleischer, T. C., Weaver, C. M., Gerbasi, V. R. & Jennings, J. L. Purifying protein complexes for mass spectrometry: applications to protein translation. *Methods* **35,** 274–290. ISSN: 10462023 (Mar. 2005).

41. Nordlund, P. *et al.* Protein production and purification. *Nature Methods* **5,** 135–146. ISSN: 1548-7105 (2008).

42. LaCava, J., Fernandez-Martinez, J., Hakhverdyan, Z. & Rout, M. P. Protein Complex Purification by Affinity Capture. *Cold Spring Harbor Protocols* **2016,** pdb.top077545. ISSN: 1940-3402 (July 2016).

43. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in Escherichia coli: advances and challenges. *Frontiers in Microbiology* **5.** ISSN: 1664-302X. doi:10.3389/fmicb.2014.00172. http://journal.frontiersin.org/article/10.3389/fmicb.2014.00172/abstract (Apr. 2014).

44. Shieh, Y.-W. *et al.* Operon structure and cotranslational subunit association direct protein assembly in bacteria. *Science (New York, N.Y.)* **350,** 678–680. ISSN: 1095-9203 (2015).

45. Poulsen, C., Holton, S., Geerlof, A., Wilmanns, M. & Song, Y.-H. Stoichiometric protein complex formation and over-expression using the prokaryotic native operon structure. *FEBS Letters* **584,** 669–674. ISSN: 00145793 (Feb. 2010).

46. Papp, B., Pál, C. & Hurst, L. D. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424,** 194–197. ISSN: 00280836 (July 2003).

47. Ni, Q. Z. *et al.* High Frequency Dynamic Nuclear Polarization. *Accounts of Chemical Research* **46,** 1933–1941. ISSN: 00014842 (ISSN) (2013).

48. Jia, B. & Jeon, C. O. High-throughput recombinant protein expression in Escherichia coli : current status and future perspectives. *Open Biology* **6,** 160196. ISSN: 2046-2441 (Aug. 2016).

49. Norimatsu, Y., Hasegawa, K., Shimizu, N. & Toyoshima, C. Protein–phospholipid interplay revealed with crystals of a calcium pump. *Nature* **545,** 193–198. ISSN: 0028-0836 (May 2017).

50. Bragg, W. H. & Bragg, W. L. The Reflection of X-rays by Crystals. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **88,** 428–438. ISSN: 1364-5021 (July 1913).

51. Shi, Y. A glimpse of structural biology through X-ray crystallography. *Cell* **159,** 995–1014. ISSN: 10974172 (2014).

52. Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature* **406,** 752–757. ISSN: 00280836 (Aug. 2000).

53. Taylor, G. The phase problem. *Acta Crystallographica Section D Biological Crystallography* **59,** 1881–1890. ISSN: 0907-4449 (Nov. 2003).

54. Robertson, J. M. An X-ray study of the phthalocyanines. Part II. Quantitative structure determination of the metal-free compound. *Journal of the Chemical Society (Resumed),* 1195–1209 (1936).

55. Dauter, Z. Use of polynuclear metal clusters in protein crystallography. *Comptes Rendus Chimie* **8,** 1808–1814. ISSN: 16310748 (2005).

56. Nozawa, K., Schneider, T. R. & Cramer, P. Core Mediator structure at 3.4 Å extends model of transcription initiation complex. *Nature* **545,** 248–251. ISSN: 0028-0836 (May 2017).

57. Hendrickson, W. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254,** 51–58. ISSN: 0036-8075 (1991).

58. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic acids research* **28,** 235–242. ISSN: 0305-1048 (2000).

59. McCoy, A. J. *et al.* Phaser crystallographic software. *Journal of Applied Crystallography* **40,** 658–674. ISSN: 0021-8898 (Aug. 2007).

60. Winn, M. D. *et al. Overview of the CCP4 suite and current developments* 2011. doi:10.1107/S0907444910045749.

61. Merk, A. *et al.* Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery. *Cell* **165,** 1698–1707. ISSN: 10974172 (2016).

62. Bai, X.-c., McMullan, G. & Scheres, S. H. How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences* **40,** 49–57. ISSN: 09680004 (Jan. 2015).

63. McMullan, G., Chen, S., Henderson, R. & Faruqi, A. Detective quantum efficiency of electron area detectors in electron microscopy. *Ultramicroscopy* **109,** 1126–1143. ISSN: 03043991 (2009).

64. Dainty, J. C. & Shaw, R. *Image Science, principles, analysis and evaluation of photographic type imaging processes* 402 (Academic Press, 1975).

65. McMullan, G., Clark, A., Turchetta, R. & Faruqi, A. Enhanced imaging in low dose electron microscopy using electron counting. *Ultramicroscopy* **109,** 1411–1416. ISSN: 03043991 (Nov. 2009).

66. Danev, R., Buijsse, B., Khoshouei, M., Plitzko, J. M. & Baumeister, W. Volta potential phase plate for in-focus phase contrast transmission electron microscopy. *Proceedings of the National Academy of Sciences* **111,** 15635–15640. ISSN: 0027-8424 (Nov. 2014).

67. Danev, R. & Baumeister, W. Cryo-EM single particle analysis with the volta phase plate. *eLife* **5,** 1–14. ISSN: 2050084X (2016).

68. Khoshouei, M., Radjainia, M., Baumeister, W. & Danev, R. Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. *Nature Communications* **8,** 16099. ISSN: 2041-1723 (June 2017).

69. Bai, X.-c., Fernandez, I. S., McMullan, G. & Scheres, S. H. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *eLife* **2.** ISSN: 2050-084X. doi:10.7554/eLife.00461. http://elifesciences.org/lookup/doi/10.7554/eLife.00461 (Feb. 2013).

70. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat Meth* **10,** 584–590. ISSN: 1548-7091 (June 2013).

71. Henderson, R. & Glaeser, R. M. Quantitative analysis of image contrast in electron micrographs of beam-sensitive crystals. *Ultramicroscopy* **16,** 139–150. ISSN: 03043991 (Jan. 1985).

72. Scheres, S. H. W. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology* **180,** 519–530. ISSN: 10478477 (2012).

73. Scheres, S. H. W. Beam-induced motion correction for sub-megadalton cryo-EM particles. *eLife* **3** (ed Kühlbrandt, W.) http://elifesciences.org/content/3/e03665.abstract (Aug. 2014).

74. Chen, S. *et al.* Structural basis for dynamic regulation of the human 26S proteasome. *Proceedings of the National Academy of Sciences of the United States of America* **113,** 12991–12996. ISSN: 1091-6490 (Nov. 2016).

75. Sigworth, F. A Maximum-Likelihood Approach to Single-Particle Image Refinement. *Journal of Structural Biology* **122,** 328–339. ISSN: 10478477 (1998).

76. Scheres, S. H. W. *et al.* Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nature Methods* **4,** 27–29. ISSN: 1548-7091 (2007).

77. Lyumkis, D., Brilot, A. F., Theobald, D. L. & Grigorieff, N. Likelihood-based classification of cryo-EM images using FREALIGN. *Journal of Structural Biology* **183,** 377–388. ISSN: 10478477 (Sept. 2013).

78. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods* **14,** 290–296. ISSN: 1548-7091 (Feb. 2017).

79. Wisedchaisri, G., Reichow, S. L. & Gonen, T. Advances in Structural and Functional Analysis of Membrane Proteins by Electron Crystallography. *Structure* **19,** 1381–1393. ISSN: 09692126 (Oct. 2011).

80. Galaz-Montoya, J. G. & Ludtke, S. J. The advent of structural biology in situ by single particle cryo-electron tomography. *Biophysics Reports* **3,** 17–35. ISSN: 2364-3439 (June 2017).

81. Bharat, T. A. M. & Scheres, S. H. W. Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in RELION. *Nature protocols* **11,** 2054–2065. ISSN: 1750-2799 (2016).

82. Leschziner, A. E. & Nogales, E. The orthogonal tilt reconstruction method: An approach to generating single-class volumes with no missing cone for ab initio reconstruction of asymmetric particles. *Journal of Structural Biology* **153,** 284–299. ISSN: 10478477 (Mar. 2006).

83. Schur, F. K. M. *et al.* An atomic model of HIV-1 capsid-SP1 reveals structures regulating assembly and maturation. *Science* **353,** 506–508. ISSN: 0036-8075 (July 2016).

84. Pervushin, K., Riek, R., Wider, G. & Wüthrich, K. Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proceedings of the National Academy of Sciences* **94,** 12366–12371. ISSN: 0027-8424 (1997).

85. Sattler, M. & Fesik, S. W. Use of deuterium labeling in NMR: overcoming a sizeable problem. *Structure* **4,** 1245–1249. ISSN: 09692126 (Nov. 1996).

86. Ollerenshaw, J. E., Tugarinov, V. & Kay, L. E. Methyl TROSY: explanation and experimental verification. *Magnetic Resonance in Chemistry* **41,** 843–852. ISSN: 0749-1581 (Oct. 2003).

87. Zhang, H. & van Ingen, H. Isotope-labeling strategies for solution NMR studies of macromolecular assemblies. *Current Opinion in Structural Biology* **38,** 75–82. ISSN: 1879033X (2016).

88. Liu, D., Xu, R. & Cowburn, D. in *Methods in enzymology* 151–175 (2009). doi:10.1016/S0076-6879(09)62008-5. http://linkinghub.elsevier.com/retrieve/pii/S0076687909620085.

89. Rosenzweig, R. *et al.* ClpB N-terminal domain plays a regulatory role in protein disaggregation. *Proceedings of the National Academy of Sciences* **112,** e6872. ISSN: 0027-8424 (2015).

90. Frederick, K. K. *et al.* Combining DNP NMR with segmental and specific labeling to study a yeast prion protein strain that is not parallel in-register. *Proceedings of the National Academy of Sciences* **114,** 3642–3647. ISSN: 0027-8424 (Apr. 2017).

91. Andrew, E. R., Bradbury, A. & Eades, R. G. Nuclear Magnetic Resonance Spectra from a Crystal rotated at High Speed. *Nature* **182,** 1659–1659. ISSN: 0028-0836 (Dec. 1958).

92. Lowe, I. J. Free Induction Decays of Rotating Solids. *Physical Review Letters* **2,** 285–287. ISSN: 0031-9007 (Apr. 1959).

93. Hansen, S. K., Bertelsen, K., Paaske, B., Nielsen, N. C. & Vosegaard, T. Solid-state NMR methods for oriented membrane proteins. *Progress in Nuclear Magnetic Resonance Spectroscopy* **88-89,** 48–85. ISSN: 00796565 (Aug. 2015).

94. Loquet, A. *et al.* Atomic model of the type III secretion system needle. *Nature* **486,** 276–9. ISSN: 1476-4687 (2012).

95. Kaplan, M. *et al.* Probing a cell-embedded megadalton protein complex by DNP-supported solid-state NMR. *Nature Methods* **12,** 5–9. ISSN: 1548-7091 (2015).

96. Huang, C. & Kalodimos, C. G. Structures of Large Protein Complexes Determined by Nuclear Magnetic Resonance Spectroscopy. *Annual Review of Biophysics* **46,** 317–336. ISSN: 1936-122X (May 2017).

97. Zeeman, P. The Effect of Magnetisation on the Nature of Light Emitted by a Substance. *Nature* **55,** 347–347. ISSN: 0028-0836 (Feb. 1897).

98. Altenbach, C., Flitsch, S. L., Khorana, H. G. & Hubbell, W. L. Structural studies on transmembrane proteins. 2. Spin labeling of bacteriorhodopsin mutants at unique cysteines. *Biochemistry* **28,** 7806–7812. ISSN: 0006-2960 (Sept. 1989).

99. Altenbach, C., Marti, T., Khorana, H. G. & Hubbell, W. L. Transmembrane protein structure: spin labeling of bacteriorhodopsin mutants. *Science (New York, N.Y.)* **248,** 1088–92. ISSN: 0036-8075 (June 1990).

100. Klare, J. P. Site-directed spin labeling EPR spectroscopy in protein research. *Biological Chemistry* **394.** ISSN: 1437-4315. doi:10.1515/hsz-2013-0155. https://www.degruyter.com/view/j/bchm.2013.394.issue-10/hsz-2013-0155/hsz-2013-0155.xml (Jan. 2013).

101. Le Breton, N. *et al.* Dimerization interface and dynamic properties of yeast IF1 revealed by Site-Directed Spin Labeling EPR spectroscopy. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1857,** 89–97. ISSN: 00052728 (Jan. 2016).

102. Jagannathan, B., Dekat, S., Golbeck, J. H. & Lakshmi, K. V. The Assembly of a Multisubunit Photosynthetic Membrane Protein Complex: A Site-Specific Spin Labeling EPR Spectroscopic Study of the PsaC Subunit in Photosystem I. *Biochemistry* **49,** 2398–2408. ISSN: 0006-2960 (Mar. 2010).

103. Levy, E. D., Boeri Erba, E., Robinson, C. V. & Teichmann, S. A. Assembly reflects evolution of protein complexes. *Nature* **453,** 1262–5. ISSN: 1476-4687 (July 2008).

104. Marsh, J. A. A. *et al.* Protein Complexes Are under Evolutionary Selection to Assemble via Ordered Pathways. en. *Cell* **153,** 461–470. ISSN: 00928674 (Apr. 2013).

105. Wan, C. *et al.* Panorama of ancient metazoan macromolecular complexes. *Nature* **525,** 339–344. ISSN: 0028-0836 (Sept. 2015).

106. Hein, M. Y. *et al.* A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell* **163,** 712–723. ISSN: 00928674 (Oct. 2015).

107. Song, F. A study of noncovalent protein complexes by matrix-assisted laser desorption/ionization. *Journal of the American Society for Mass Spectrometry* **18,** 1286–1290. ISSN: 1044-0305 (July 2007).

108. Krusemark, C. J., Frey, B. L., Belshaw, P. J. & Smith, L. M. Modifying the charge state distribution of proteins in electrospray ionization mass spectrometry by chemical derivatization. *Journal of the American Society for Mass Spectrometry* **20,** 1617–1625. ISSN: 1044-0305 (Sept. 2009).

109. Radionova, A., Filippov, I. & Derrick, P. J. In pursuit of resolution in time-of-flight mass spectrometry: A historical perspective. *Mass spectrometry reviews* **35,** 738–757. ISSN: 1098-2787 (Oct. 2016).

110. Hu, Q. *et al.* The Orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry* **40,** 430–443. ISSN: 1076-5174 (Apr. 2005).

111. Wilm, M. S. & Mann, M. Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? *International Journal of Mass Spectrometry and Ion Processes* **136,** 167–180. ISSN: 01681176 (Sept. 1994).

112. El-Faramawy, A., Siu, K. W. M. & Thomson, B. A. Efficiency of nano-electrospray ionization. *Journal of the American Society for Mass Spectrometry* **16,** 1702–1707. ISSN: 1044-0305 (Oct. 2005).

113. Sobott, F., Hernández, H., McCammon, M. G., Tito, M. A. & Robinson, C. V. A tandem mass spectrometer for improved transmission and analysis of large macromolecular assemblies. *Analytical Chemistry* **74,** 1402–1407. ISSN: 00032700 (2002).

114. Hernandez, H. & Robinson, C. V. Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nat. Protoc.* **2,** 715–726. ISSN: 1750-2799 (2007).

115. Sobott, F., Benesch, J. L. P., Vierling, E. & Robinson, C. V. Subunit Exchange of Multimeric Protein Complexes. *Journal of Biological Chemistry* **277,** 38921–38929. ISSN: 0021-9258 (Oct. 2002).

116. Laganowsky, A. *et al.* Membrane proteins bind lipids selectively to modulate their structure and function. *Nature* **510,** 172–175. ISSN: 0028-0836 (June 2014).

117. Stengel, F., Aebersold, R. & Robinson, C. V. Joining Forces: Integrating Proteomics and Cross-linking with the Mass Spectrometry of Intact Complexes. *Molecular & Cellular Proteomics* **11,** R111.014027–R111.014027. ISSN: 1535-9476 (Mar. 2012).

118. Ward, A. B., Sali, A. & Wilson, I. A. Integrative Structural Biology. *Science* **339,** 913–915. ISSN: 0036-8075 (Feb. 2013).

119. Van den Bedem, H. & Fraser, J. S. Integrative, dynamic structural biology at atomic resolution - it's about time. *Nature Methods* **12,** 307–318. ISSN: 1548-7091 (Mar. 2015).

120. Leitner, A., Faini, M., Stengel, F. & Aebersold, R. Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends in Biochemical Sciences* **41,** 20–32. ISSN: 13624326 (2016).

121. Suchanek, M., Radzikowska, A. & Thiele, C. Photo-leucine and photo-methionine allow identification of protein-protein interactions in living cells. *Nature Methods* **2,** 261–268. ISSN: 1548-7091 (Apr. 2005).

122. Barysz, H. *et al.* Three-dimensional topology of the SMC2/SMC4 subcomplex from chicken condensin I revealed by cross-linking and molecular modelling. *Open biology* **5,** 150005. ISSN: 2046-2441 (2015).

123. Beck, M. & Hurt, E. The nuclear pore complex: understanding its function through structural insight. *Nature Reviews Molecular Cell Biology.* ISSN: 1471-0072. doi:10.1038/nrm.2016.147. http://dx.doi.org/10.1038/nrm.2016.147%20http://www.nature.com/doifinder/10.1038/nrm.2016.147 (Dec. 2016).

124. Bui, K. H. *et al.* Integrated Structural Analysis of the Human Nuclear Pore Complex Scaffold. *Cell* **155,** 1233–1243. ISSN: 00928674 (Dec. 2013).

125. Shi, Y. *et al.* Structural Characterization by Cross-linking Reveals the Detailed Architecture of a Coatomer-related Heptameric Module from the Nuclear Pore Complex. *Molecular & Cellular Proteomics* **13,** 2927–2943. ISSN: 1535-9476 (Nov. 2014).

126. Oeffinger, M. Two steps forward-one step back: Advances in affinity purification mass spectrometry of macromolecular complexes. *Proteomics* **12,** 1591–1608. ISSN: 16159853 (May 2012).

127. Morris, J. H. *et al.* Affinity purification–mass spectrometry and network analysis to understand protein-protein interactions. *Nature Protocols* **9,** 2539–2554. ISSN: 1754-2189 (2014).

128. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537,** 347–355. ISSN: 0028-0836 (2016).

129. Malovannaya, A. *et al.* Analysis of the Human Endogenous Coregulator Complexome. *Cell* **145,** 787–799. ISSN: 00928674 (May 2011).

130. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162,** 425–440. ISSN: 00928674 (2015).

131. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology* **17,** 1030–1032. ISSN: 10870156 (Oct. 1999).

132. Hubner, N. C. *et al.* Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *The Journal of Cell Biology* **189,** 739–754. ISSN: 0021-9525 (May 2010).

133. Selbach, M. & Mann, M. Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK). *Nature Methods* **3,** 981–983. ISSN: 1548-7091 (Dec. 2006).

134. Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G. & Orengo, C. Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. *Structure* **18,** 1233–1243. ISSN: 09692126 (Oct. 2010).

135. Keilhauer, E. C., Hein, M. Y. & Mann, M. Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Molecular & cellular proteomics : MCP* **14,** 120–35. ISSN: 1535-9484 (2015).

136. Ong, S.-E. *et al.* Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Molecular & Cellular Proteomics* **1.5,** 376–386. ISSN: 15359476 (2002).

137. Ross, P. L. Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Molecular & Cellular Proteomics* **3,** 1154–1169. ISSN: 1535-9476 (Sept. 2004).

138. Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology* **17,** 994–999. ISSN: 10870156 (Oct. 1999).

139. Thompson, A. *et al.* Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Analytical Chemistry* **75,** 1895–1904. ISSN: 0003-2700 (Apr. 2003).

140. Liu, H., Sadygov, R. G. & Yates, J. R. A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Analytical Chemistry* **76,** 4193–4201. ISSN: 0003-2700 (July 2004).

141. Zybailov, B., Coleman, M. K., Florens, L. & Washburn, M. P. Correlation of Relative Abundance Ratios Derived from Peptide Ion Chromatograms and Spectrum Counting for Quantitative Proteomic Analysis Using Stable Isotope Labeling. *Analytical Chemistry* **77,** 6218–6224. ISSN: 0003-2700 (Oct. 2005).

142. Lundgren, D. H., Hwang, S.-I., Wu, L. & Han, D. K. Role of spectral counting in quantitative proteomics. *Expert Review of Proteomics* **7,** 39–53. ISSN: 1478-9450 (Feb. 2010).

143. Nahnsen, S., Bielow, C., Reinert, K. & Kohlbacher, O. Tools for Label-free Peptide Quantification. *Molecular & Cellular Proteomics* **12,** 549–556. ISSN: 1535-9476 (2013).

144. Fabre, B. *et al.* Comparison of label-free quantification methods for the determination of protein complexes subunits stoichiometry. *EuPA Open Proteomics* **4,** 82–86. ISSN: 22129685 (Sept. 2014).

145. Cox, J. *et al.* Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Molecular & Cellular Proteomics* **13,** 2513–2526. ISSN: 1535-9476 (Sept. 2014).

146. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26,** 1367–1372. ISSN: 1087-0156 (Dec. 2008).

147. Schaeffer, S. E. Graph clustering. *Computer Science Review* **1,** 27–64. ISSN: 15740137 (Aug. 2007).

148. Van Dongen, S. A Cluster algorithm for graphs. eng. *Report - Information systems,* 1–40. ISSN: 1386-3681 (2000).

149. Li, X., Wu, M., Kwoh, C.-K. & Ng, S.-K. Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Genomics* **11,** S3. ISSN: 1471-2164 (2010).

150. Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440,** 637–643. ISSN: 0028-0836 (Mar. 2006).

151. Wu, M., Li, X., Kwoh, C.-K. & Ng, S.-K. A core-attachment based method to detect protein complexes in PPI networks. *BMC Bioinformatics* **10,** 169. ISSN: 1471-2105 (2009).

152. Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods* **9,** 471–472. ISSN: 1548-7091 (Mar. 2012).

153. Montaño-Gutierrez, L. F., Ohta, S., Kustatscher, G., Earnshaw, W. C. & Rappsilber, J. Nano Random Forests to mine protein complexes and their relationships in quantitative proteomics data. *Molecular biology of the cell* **28,** 673–680. ISSN: 1939-4586 (2017).

154. Abbe, E. Beiträge zur Theorie des Mikroskops und der mikroskopischen Wahrnehmung. *Archiv für Mikroskopische Anatomie* **9,** 413–418. ISSN: 0176-7364 (1873).

155. Hell, S. W. & Wichmann, J. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Optics Letters* **19,** 780. ISSN: 0146-9592 (June 1994).

156. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods* **3,** 793–796. ISSN: 1548-7091 (Oct. 2006).

157. Hess, S. T., Girirajan, T. P. & Mason, M. D. Ultra-High Resolution Imaging by Fluorescence Photoactivation Localization Microscopy. *Biophysical Journal* **91,** 4258–4272. ISSN: 00063495 (Dec. 2006).

158. Betzig, E. *et al.* Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. *Science* **313,** 1642–1645. ISSN: 0036-8075 (Sept. 2006).

159. Schrödinger, E. Are there quantum jumps? (Pt. II). *The British Journal for the Philosophy of Science* **III,** 233–242. ISSN: 0007-0882 (1952).

160. Szymborska, A. *et al.* Nuclear Pore Scaffold Structure Analyzed by Super-Resolution Microscopy and Particle Averaging. *Science* **341,** 655–658. ISSN: 0036-8075 (Aug. 2013).

161. Ries, J., Kaplan, C., Platonova, E., Eghlidi, H. & Ewers, H. A simple, versatile method for GFP-based super-resolution microscopy via nanobodies. *Nature Methods* **9,** 582–584. ISSN: 1548-7091 (Apr. 2012).

162. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **181,** 223–230. ISSN: 0036-8075 (1973).

163. Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Genetics* **23,** ii–iv. ISSN: 0887-3585 (Nov. 1995).

164. Janin, J. *et al.* CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins: Structure, Function, and Genetics* **52,** 2–9. ISSN: 0887-3585 (July 2003).

165. Haas, J. *et al.* The Protein Model Portal–a comprehensive resource for protein structure and model information. *Database* **2013,** bat031–bat031. ISSN: 1758-0463 (Apr. 2013).

166. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins: Structure, Function, and Bioinformatics* **84,** 4–14. ISSN: 08873585 (Sept. 2016).

167. Jiang, Z.-Y. *et al.* Insight into the Intermolecular Recognition Mechanism between Keap1 and IKKβ Combining Homology Modelling, Protein-Protein Docking, Molecular Dynamics Simulations and Virtual Alanine Mutation. *PLoS ONE* **8** (ed Kleinjung, J.) e75076. ISSN: 1932-6203 (Sept. 2013).

168. Rajapaksha, H. & Petrovsky, N. In Silico Structural Homology Modelling and Docking for Assessment of Pandemic Potential of a Novel H7N9 Influenza Virus and Its Ability to Be Neutralized by Existing Anti-Hemagglutinin Antibodies. *PLoS ONE* **9** (ed Guan, Y.) e102618. ISSN: 1932-6203 (July 2014).

169. Agostino, M., Mancera, R. L., Ramsland, P. A. & Fernández-Recio, J. Optimization of protein-protein docking for predicting Fc-protein interactions. *Journal of Molecular Recognition* **29,** 555–568. ISSN: 09523499 (Nov. 2016).

170. Lensink, M. F. *et al.* Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics* **84,** 323–348. ISSN: 08873585 (Sept. 2016).

171. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *The EMBO journal.* ISSN: 0261-4189. doi:060fehlt (1986).

172. Chen, H. & Skolnick, J. M-TASSER: An Algorithm for Protein Quaternary Structure Prediction. *Biophysical Journal* **94,** 918–928. ISSN: 00063495 (2008).

173. Tuncbag, N., Gursoy, A., Nussinov, R. & Keskin, O. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nature Protocols* **6,** 1341–1354. ISSN: 1754-2189 (2011).

174. Guerler, A., Govindarajoo, B. & Zhang, Y. Mapping monomeric threading to protein-protein structure prediction. *Journal of Chemical Information and Modeling* **53,** 717–725. ISSN: 15499596 (2013).

175. Bowie, J., Luthy, R. & Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253,** 164–170. ISSN: 0036-8075 (1991).

176. Lu, L., Lu, H. & Skolnick, J. Multiprospector: An algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins: Structure, Function and Genetics* **49,** 350–364. ISSN: 08873585 (2002).

177. Szilagyi, A. & Zhang, Y. Template-based structure modeling of protein-protein interactions. *Current Opinion in Structural Biology* **24,** 10–23. ISSN: 0959440X (2014).

178. Huang, S.-Y. Search strategies and evaluation in protein?protein docking: principles, advances and challenges. *Drug Discovery Today* **19,** 1081–1096. ISSN: 13596446 (Aug. 2014).

179. Katchalski-Katzir, E. *et al.* Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences* **89,** 2195–2199. ISSN: 0027-8424 (Mar. 1992).

180. Hart, T. N. & Read, R. J. A multiple-start Monte Carlo docking method. *Proteins: Structure, Function, and Genetics* **13,** 206–222. ISSN: 0887-3585 (July 1992).

181. Zacharias, M. ATTRACT: Protein-protein docking in CAPRI using a reduced protein model. *Proteins: Structure, Function, and Bioinformatics* **60,** 252–256. ISSN: 08873585 (June 2005).

182. Lyskov, S. & Gray, J. J. The RosettaDock server for local protein-protein docking. *Nucleic Acids Research* **36,** W233–W238. ISSN: 0305-1048 (May 2008).

183. Zhang, Z., Schindler, C. E. M., Lange, O. F. & Zacharias, M. Application of Enhanced Sampling Monte Carlo Methods for High-Resolution Protein-Protein Docking in Rosetta. *PLOS ONE* **10** (ed Colombo, G.) e0125941. ISSN: 1932-6203 (June 2015).

184. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information. *Journal of the American Chemical Society* **125,** 1731–1737. ISSN: 0002-7863 (Feb. 2003).

185. Van Zundert, G. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology* **428,** 720–725. ISSN: 00222836 (Feb. 2016).

186. Kynast, P., Derreumaux, P. & Strodel, B. Evaluation of the coarse-grained OPEP force field for protein-protein docking. *BMC Biophysics* **9,** 4. ISSN: 2046-1682 (Dec. 2016).

187. Böhm, H.-J. Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *Journal of Computer-Aided Molecular Design* **12,** 309–309. ISSN: 0920654X (1998).

188. Sasse, A., de Vries, S. J., Schindler, C. E. M., de Beauchêne, I. C. & Zacharias, M. Rapid Design of Knowledge-Based Scoring Potentials for Enrichment of Near-Native Geometries in Protein-Protein Docking. *PLOS ONE* **12** (ed Sticht, H.) e0170625. ISSN: 1932-6203 (Jan. 2017).

189. Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Research* **43,** W389–W394. ISSN: 0305-1048 (July 2015).

190. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* **44,** D279–D285. ISSN: 0305-1048 (Jan. 2016).

191. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45,** D158–D169. ISSN: 0305-1048 (Jan. 2017).

192. Altschuh, D., Lesk, A. M., Bloomer, A. C. & Klug, A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of Molecular Biology* **193,** 693–707. ISSN: 00222836 (1987).

193. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences* **106,** 67–72. ISSN: 0027-8424 (Jan. 2009).

194. Lunt, B. *et al.* Inference of direct residue contacts in two-component signaling. *Methods in enzymology* **471,** 17–41. ISSN: 1557-7988 (2010).

195. Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* **27,** 379–423. ISSN: 15387305 (1948).

196. Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6.** ISSN: 19326203. doi:10.1371/journal.pone.0028766. arXiv: 1110.5091 (2011).

197. Marks, D. S., Hopf, T. a. & Sander, C. Protein structure prediction from sequence variation. *Nature biotechnology* **30,** 1072–80. ISSN: 1546-1696 (Nov. 2012).

198. Hopf, T. a. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **3,** 1–45. ISSN: 2050-084X (Oct. 2014).

199. Wagner, A. The Yeast Protein Interaction Network Evolves Rapidly and Contains Few Redundant Duplicate Genes. *Molecular Biology and Evolution* **18,** 1283–1292. ISSN: 0737-4038 (July 2001).

200. Wagner, A. How the global structure of protein interaction networks evolves. *Proceedings of the Royal Society B: Biological Sciences* **270,** 457–466. ISSN: 0962-8452 (2003).

201. Fokkens, L., Hogeweg, P. & Snel, B. Gene duplications contribute to the overrepresentation of interactions between proteins of a similar age. *BMC Evolutionary Biology* **12,** 99. ISSN: 1471-2148 (2012).

202. Brum, J. R. *et al.* Illuminating structural proteins in viral "dark matter" with metaproteomics. *Proceedings of the National Academy of Sciences* **113,** 2436–2441. ISSN: 0027-8424 (Mar. 2016).

203. Shi, M. *et al.* Redefining the invertebrate RNA virosphere. *Nature* **540,** 539–543. ISSN: 0028-0836 (Nov. 2016).

204. Mukherjee, S. *et al.* 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nature Biotechnology.* ISSN: 1087-0156. doi:10.1038/nbt.3886. http://www.nature.com/doifinder/10.1038/nbt.3886 (June 2017).

205. Hu, H. & Sun, Y. Molecular dynamics simulations of disjoining pressure effect in ultra-thin water film on a metal surface. *Applied Physics Letters* **103,** 263110. ISSN: 0003-6951 (Dec. 2013).

206. Rapaport, D. C. Molecular dynamics simulation: a tool for exploration and discovery using simple models. *Journal of Physics: Condensed Matter* **26,** 503104. ISSN: 0953-8984 (Dec. 2014).

207. Shao, Y. *et al.* Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Molecular Physics* **113,** 184–215. ISSN: 0026-8976 (Jan. 2015).

208. Buch, I., Giorgino, T. & De Fabritiis, G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences* **108,** 10184–10189. ISSN: 0027-8424 (June 2011).

209. Zhao, G. *et al.* Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* **497,** 643–646. ISSN: 0028-0836 (May 2013).

210. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Research* **33,** W382–W388. ISSN: 0305-1048 (July 2005).

211. Plattner, N., Doerr, S., De Fabritiis, G. & Noé, F. Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nature Chemistry,* 1–7. ISSN: 1755-4330 (2017).

212. Orchard, S. *et al.* The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research* **42,** D358–63. ISSN: 1362-4962 (Jan. 2014).

213. Meldal, B. H. M. *et al.* The complex portal - an encyclopaedia of macromolecular complexes. *Nucleic Acids Research* **43,** D479–D484. ISSN: 0305-1048 (Jan. 2015).

214. Drew, K. *et al.* Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular Systems Biology* **13,** 932. ISSN: 1744-4292 (June 2017).

215. Ruepp, A. *et al.* CORUM: The comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Research* **38,** 497–501. ISSN: 03051048 (2009).

216. Havugimana, P. C. *et al.* A Census of Human Soluble Protein Complexes. en. *Cell* **150,** 1068–1081. ISSN: 00928674 (Aug. 2012).

217. Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545,** 505–509. ISSN: 0028-0836 (May 2017).

218. Lam, Y. W., Lamond, A. I., Mann, M. & Andersen, J. S. Analysis of Nucleolar Protein Dynamics Reveals the Nuclear Degradation of Ribosomal Proteins. *Current Biology* **17,** 749–760. ISSN: 09609822 (2007).

219. Sung, M.-K., Reitsma, J. M., Sweredoski, M. J., Hess, S. & Deshaies, R. J. Ribosomal proteins produced in excess are degraded by the ubiquitin-proteasome system. *Molecular Biology of the Cell* **27,** 2642–2652. ISSN: 1059-1524 (Sept. 2016).

220. Yang, J.-R., Liao, B.-Y., Zhuang, S.-M. & Zhang, J. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America* **109,** E831–40. ISSN: 1091-6490 (Apr. 2012).

221. Blikstad, I., James Nelson, W., Moon, R. T. & Lazarides, E. Synthesis and assembly of spectrin during avian erythropoiesis: Stoichiometric assembly but unequal synthesis of α and β spectrin. *Cell* **32,** 1081–1091. ISSN: 00928674 (Apr. 1983).

222. Hanspal, M. Synthesis and assembly of membrane skeletal proteins in mammalian red cell precursors. *The Journal of Cell Biology* **105,** 1417–1424. ISSN: 0021-9525 (Sept. 1987).

223. Lehnert, M. E. & Lodish, H. F. Unequal synthesis and differential degradation of alpha and beta spectrin during murine erythroid differentiation. *The Journal of cell biology* **107,** 413–26. ISSN: 0021-9525 (Aug. 1988).

224. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–42. ISSN: 1476-4687 (May 2011).

225. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324,** 218–223. ISSN: 0036-8075 (Apr. 2009).

226. Ingolia, N. T. Ribosome Footprint Profiling of Translation throughout the Genome. *Cell* **165,** 22–33. ISSN: 10974172 (2016).

227. Li, G. W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157,** 624–635. ISSN: 10974172 (2014).

228. Brandman, O. *et al.* A Ribosome-Bound Quality Control Complex Triggers Degradation of Nascent Peptides and Signals Translation Stress. *Cell* **151,** 1042–1054. ISSN: 00928674 (Nov. 2012).

229. Mallik, S. & Kundu, S. Coevolutionary constraints in the sequence-space of macromolecular complexes reflect their self-assembly pathways. *Proteins: Structure, Function, and Bioinformatics* **85,** 1183–1189. ISSN: 08873585 (July 2017).

230. Duncan, C. D. & Mata, J. Widespread cotranslational formation of protein complexes. *PLoS genetics* **7,** e1002398 (2011).

231. Mertens, H. D. & Svergun, D. I. Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *Journal of Structural Biology* **172,** 128–141. ISSN: 10478477 (Oct. 2010).

232. Zhang, Z. & Vachet, R. W. Kinetics of Protein Complex Dissociation Studied by Hydrogen/Deuterium Exchange and Mass Spectrometry. *Analytical Chemistry* **87,** 11777–11783. ISSN: 0003-2700 (Dec. 2015).

233. Bernecky, C., Herzog, F., Baumeister, W., Plitzko, J. M. & Cramer, P. Structure of transcribing mammalian RNA polymerase II. *Nature* **529,** 551–554. ISSN: 0028-0836 (Jan. 2016).

234. Fernandez-Martinez, J. *et al.* Structure and Function of the Nuclear Pore Complex Cytoplasmic mRNA Export Platform. *Cell* **167,** 1215–1228.e25. ISSN: 00928674 (Nov. 2016).

235. Tsai, K.-L. *et al.* Mediator structure and rearrangements required for holoenzyme formation. *Nature* **544,** 196–201. ISSN: 0028-0836 (Mar. 2017).

236. Cassiday, L. Structural biology: More than a crystallographer. *Nature* **505,** 711–713. ISSN: 0028-0836 (Jan. 2014).

237. Appolaire, A. *et al.* Small-angle neutron scattering reveals the assembly mode and oligomeric architecture of TET, a large, dodecameric aminopeptidase. *Acta Crystallographica Section D Biological Crystallography* **70,** 2983–2993. ISSN: 1399-0047 (Nov. 2014).

238. Macek, P. *et al.* Unraveling self-assembly pathways of the 468-kDa proteolytic machine TET2, 1–10 (2017).

239. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–81. ISSN: 1476-4687 (May 2014).

240. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–7. ISSN: 1476-4687 (May 2014).

241. Ezkurdia, I., V?zquez, J., Valencia, A. & Tress, M. Analyzing the First Drafts of the Human Proteome. *Journal of Proteome Research* **13,** 3854–3855. ISSN: 1535-3893 (Aug. 2014).

242. Macaulay, I. C., Ponting, C. P. & Voet, T. Single-Cell Multiomics: Multiple Measurements from Single Cells. *Trends in Genetics* **33,** 155–168. ISSN: 01689525 (Feb. 2017).

243. Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences* **99,** 12795–12800. ISSN: 0027-8424 (2002).

244. Ross, C. A. & Poirier, M. A. Protein aggregation and neurodegenerative disease. *Nature Medicine* **10,** S10–S17. ISSN: 1078-8956 (July 2004).

245. Mushegian, A. R. & Koonin, E. V. Gene order is not conserved in bacterial evolution. *Trends in Genetics* **12,** 289–290. ISSN: 01689525 (Aug. 1996).

246. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order : a fingerprint of proteins that physically interact. *Trends in biochemical sciences* **0004,** 324–328. ISSN: 0968-0004 (Sept. 1998).

247. Swain, P. S. Efficient attenuation of stochasticity in gene expression through post-transcriptional control. *Journal of Molecular Biology* **344,** 965–976. ISSN: 00222836 (2004).

248. Sneppen, K., Pedersen, S., Krishna, S., Dodd, I. & Semsey, S. Economy of operon formation: Cotranscription minimizes shortfall in protein complexes. *mBio* **1,** 3–5. ISSN: 21507511 (2010).

249. Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D. & von Mering, C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* **15,** 3163–3168. ISSN: 16159853 (2015).

250. Kovács, K., Hurst, L. D. & Papp, B. Stochasticity in protein levels drives colinearity of gene order in metabolic operons of Escherichia coli. *PLoS biology* **7,** e1000115. ISSN: 1545-7885 (May 2009).

251. Pasek, S., Risler, J.-L. & Brezellec, P. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics* **22,** 1418–1423. ISSN: 1367-4803 (June 2006).

252. Byrne, R., Levin, J. G., Bladen, H. a. & Nirenberg, M. W. The in vitro formation of a DNA-ribosome complex. *Proceedings of the National Academy of Sciences* **52,** 140–148. ISSN: 0027-8424 (July 1964).

253. Gowrishankar, J. & Harinarayanan, R. Why is transcription coupled to translation in bacteria? *Molecular Microbiology* **54,** 598–603. ISSN: 0950382X (Sept. 2004).

254. Kohler, R., Mooney, R. A., Mills, D. J., Landick, R. & Cramer, P. Architecture of a transcribing-translating expressome. *Science* **356,** 194–197. ISSN: 0036-8075 (Apr. 2017).

255. Oppenheim, D. S. & Yanofsky, C. Translational coupling during expression of the tryptophan operon of Escherichia coli. *Genetics* **95,** 785–795. ISSN: 00166731 (1980).

256. Levin-Karp, A. *et al.* Quantifying Translational Coupling in E. coli Synthetic Operons Using RBS Modulation and Fluorescent Reporters. *ACS Synthetic Biology* **2,** 327–336. ISSN: 2161-5063 (June 2013).

257. Nishizaki, T., Tsuge, K., Itaya, M., Doi, N. & Yanagawa, H. Metabolic engineering of carotenoid biosynthesis in Escherichia coli by ordered gene assembly in Bacillus subtilis. *Applied and Environmental Microbiology* **73,** 1355–1361. ISSN: 00992240 (2007).

258. Lim, H. N., Lee, Y. & Hussein, R. Fundamental relationship between operon organization and gene expression. *Proceedings of the National Academy of Sciences* **108,** 10626–31. ISSN: 1091-6490 (June 2011).

259. Zaslaver, A. *et al.* Just-in-time transcription program in metabolic pathways. *Nature genetics* **36,** 486–91. ISSN: 1061-4036 (May 2004).

260. Huntley, R. P. *et al.* The GOA database: gene Ontology annotation updates for 2015. *Nucleic acids research* **43,** D1057–63. ISSN: 1362-4962 (Jan. 2015).

261. Ellis, R. J. Molecular chaperones: assisting assembly in addition to folding. *Trends in Biochemical Sciences* **31,** 395–401. ISSN: 09680004 (2006).

262. Koide, T. *et al.* Prevalence of transcription promoters within archaeal operons and coding sequences. *Molecular Systems Biology* **5.** ISSN: 1744-4292. doi:10.1038/msb.2009.42. http://msb.embopress.org/cgi/doi/10.1038/msb.2009.42 (June 2009).

263. Conway, T. *et al.* Unprecedented High-Resolution View of Bacterial Operon Architecture Revealed by RNA Sequencing. *mBio* **5,** e01442–14–e01442–14. ISSN: 2150-7511 (July 2014).

264. Wells, J. N., Bergendahl, L. T. & Marsh, J. A. Co-translational assembly of protein complexes. *Biochemical Society Transactions* **43,** 1221–1226. ISSN: 0300-5127 (Dec. 2015).

265. Natan, E., Wells, J. N., Teichmann, S. A. & Marsh, J. A. Regulation, evolution and consequences of cotranslational protein complex assembly. *Current Opinion in Structural Biology* **42,** 90–97. ISSN: 0959-440X (Feb. 2017).

266. Nuñez, P. a., Romero, H., Farber, M. D. & Rocha, E. P. C. Natural selection for operons depends on genome size. *Genome biology and evolution* **5,** 2242–54. ISSN: 1759-6653 (Jan. 2013).

267. Burkhardt, D. H. *et al.* Operon mRNAs are organized into ORF-centric structures that predict translation efficiency. *eLife* **6.** ISSN: 2050-084X. doi:10.7554/eLife.22037. http://elifesciences.org/lookup/doi/10.7554/eLife.22037 (Jan. 2017).

268. Warner, J. R. The economics of ribosome biosynthesis in yeast. *Trends in Biochemical Sciences* **24,** 437–440. ISSN: 09680004 (Nov. 1999).

269. Lehner, B. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Molecular Systems Biology* **4.** ISSN: 1744-4292. doi:10.1038/msb.2008.11. http://msb.embopress.org/cgi/doi/10.1038/msb.2008.11 (Mar. 2008).

270. Schuster-Böckler, B., Conrad, D. & Bateman, A. Dosage sensitivity shapes the evolution of copy-number varied regions. *PLoS ONE* **5,** 1–10. ISSN: 19326203 (2010).

271. Schimke, R. T. & Doyle, D. Control of Enzyme Levels in Animal Tissues. *Annual Review of Biochemistry* **39,** 929–976. ISSN: 0066-4154 (June 1970).

272. Goldberg, a. L. & Dice, J. F. Intracellular protein degradation in mammalian and bacterial cells. *Annual review of biochemistry* **43,** 835–869. ISSN: 0066-4154 (1974).

273. Wheatley, D. N., Giddings, M. R. & Inglis, M. S. Kinetics of degradation of "short-" and "long-lived" proteins in cultured mammalian cells. *Cell biology international reports* **4,** 1081–90. ISSN: 0309-1651 (Dec. 1980).

274. Tyler, R. E. *et al.* Unassembled CD147 is an endogenous endoplasmic reticulum-associated degradation substrate. *Molecular Biology of the Cell* **23,** 4668–4678. ISSN: 1059-1524 (Dec. 2012).

275. Ward, C. L. & Kopito, R. R. Intracellular turnover of cystic fibrosis transmembrane conductance regulator. Inefficient processing and rapid degradation of wild-type and mutant proteins. *The Journal of biological chemistry* **269,** 25710–8. ISSN: 0021-9258 (Oct. 1994).

276. Kim, W. *et al.* Systematic and Quantitative Assessment of the Ubiquitin-Modified Proteome. *Molecular Cell* **44,** 325–340. ISSN: 10972765 (Oct. 2011).

277. Wang, F., Durfee, L. A. & Huibregtse, J. M. A Cotranslational Ubiquitination Pathway for Quality Control of Misfolded Proteins. *Molecular Cell* **50,** 368–378. ISSN: 10972765 (May 2013).

278. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* **147,** 789–802. ISSN: 00928674 (Nov. 2011).

279. Doherty, M. K., Hammond, D. E., Clague, M. J., Gaskell, S. J. & Beynon, R. J. Turnover of the Human Proteome: Determination of Protein Intracellular Stability by Dynamic SILAC. *Journal of Proteome Research* **8,** 104–112. ISSN: 1535-3893 (Jan. 2009).

280. Kristensen, A. R., Gsponer, J. & Foster, L. J. Protein synthesis rate is the predominant regulator of protein expression during differentiation. *Molecular Systems Biology* **9,** 689–689. ISSN: 1744-4292 (Apr. 2013).

281. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Molecular and cellular biology* **19,** 1720–30. ISSN: 0270-7306 (1999).

282. Chen, G. *et al.* Discordant Protein and mRNA Expression in Lung Adenocarcinomas. *Molecular & Cellular Proteomics* **1,** 304–313. ISSN: 1535-9476 (Apr. 2002).

283. Kiick, K. L., Saxon, E., Tirrell, D. A. & Bertozzi, C. R. Incorporation of azides into recombinant proteins for chemoselective modification by the Staudinger ligation. *Proceedings of the National Academy of Sciences* **99,** 19–24. ISSN: 0027-8424 (Jan. 2002).

284. Dieterich, D. C., Link, A. J., Graumann, J., Tirrell, D. A. & Schuman, E. M. Selective identification of newly synthesized proteins in mammalian cells using bioorthogonal noncanonical amino acid tagging (BONCAT). *Proceedings of the National Academy of Sciences* **103,** 9482–9487. ISSN: 0027-8424 (June 2006).

285. Larance, M., Ahmad, Y., Kirkwood, K. J., Ly, T. & Lamond, A. I. Global Subcellular Characterization of Protein Degradation Using Quantitative Proteomics. *Molecular & Cellular Proteomics* **12,** 638–650. ISSN: 1535-9476 (Mar. 2013).

286. Eichelbaum, K. & Krijgsveld, J. Rapid Temporal Dynamics of Transcription, Protein Synthesis, and Secretion during Macrophage Activation. *Molecular & Cellular Proteomics* **13,** 792–810. ISSN: 1535-9476 (Mar. 2014).

287. Deneke, C., Lipowsky, R. & Valleriani, A. Complex Degradation Processes Lead to Non-Exponential Decay Patterns and Age-Dependent Decay Rates of Messenger RNA. *PLoS ONE* **8** (ed Tsimring, L.) e55442. ISSN: 1932-6203 (Feb. 2013).

288. Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19,** 716–723. ISSN: 0018-9286 (Dec. 1974).

289. Ori, A. *et al.* Spatiotemporal variation of mammalian protein complex stoichiometries. *Genome Biology* **17,** 47. ISSN: 1474-760X (Dec. 2016).

290. Perica, T. *et al.* The emergence of protein complexes: quaternary structure, dynamics and allostery. en. *Biochemical Society Transactions* **40,** 475–491. ISSN: 0300-5127, 1470-8752 (June 2012).

291. Marsh, J. A. & Teichmann, S. A. Structure, Dynamics, Assembly, and Evolution of Protein Complexes. *Annual Review of Biochemistry* **84,** 141210135300003. ISSN: 0066-4154 (2015).

292. Goldberg, A. L. Protein degradation and protection against misfolded or damaged proteins. *Nature* **426,** 895–9. ISSN: 1476-4687 (Dec. 2003).

293. Malinverni, J. C. *et al.* YfiO stabilizes the YaeT complex and is essential for outer membrane protein assembly in Escherichia coli. *Molecular Microbiology* **61,** 151–164. ISSN: 0950382X (July 2006).

294. Toyama, B. H. *et al.* Identification of long-lived proteins reveals exceptional stability of essential cellular structures. *Cell* **154,** 971–982. ISSN: 1097-4172 (Aug. 2013).

295. Arnold, T. B. & Emerson, J. W. Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions. *The R Journal* **3,** 34–39 (2011).

296. Leibiger, C. *et al.* First Molecular Cytogenetic High Resolution Characterization of the NIH 3T3 Cell Line by Murine Multicolor Banding. *Journal of Histochemistry & Cytochemistry* **61,** 306–312. ISSN: 0022-1554 (Apr. 2013).

297. Brown, A. A. *et al.* Local genetic effects on gene expression across 44 human tissues. *bioRxiv* (2016).

298. Chen, T. *et al.* mUbiSiDa: A Comprehensive Database for Protein Ubiquitination Sites in Mammals. *PLoS ONE* **9** (ed Raghava, G. P. S.) e85744. ISSN: 1932-6203 (Jan. 2014).

299. Hose, J. *et al.* Dosage compensation can buffer copynumber variation in wild yeast. *eLife* **4.** ISSN: 2050084X. doi:10.7554/eLife.05462 (2015).

300. Presson, A. P. *et al.* Current estimate of Down Syndrome population prevalence in the United States. *The Journal of pediatrics* **163,** 1163–8. ISSN: 1097-6833 (Oct. 2013).

301. Jia, C.-W. *et al.* Aneuploidy in Early Miscarriage and its Related Factors. *Chinese Medical Journal* **128,** 2772. ISSN: 0366-6999 (2015).

302. Lightfoot, D. A., Kouznetsova, A., Mahdy, E., Wilbertz, J. & Höög, C. The fate of mosaic aneuploid embryos during mouse development. *Developmental Biology* **289,** 384–394. ISSN: 00121606 (Jan. 2006).

303. Compton, D. A. Mechanisms of aneuploidy. *Current Opinion in Cell Biology* **23,** 109–113. ISSN: 09550674 (Feb. 2011).

304. Stingele, S. *et al.* Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Molecular Systems Biology* **8.** ISSN: 1744-4292. doi:10.1038/msb.2012.40. http://msb.embopress.org/cgi/doi/10.1038/msb.2012.40 (Sept. 2012).

305. Dephoure, N. *et al.* Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. *eLife,* e03023. ISSN: 2050-084X (2014).

306. Gonçalves, E. *et al.* Chromosomal rearrangements are commonly post-transcriptionally attenuated in cancer. *bioRxiv,* 1–40 (2017).

307. Pu, S., Wong, J., Turner, B., Cho, E. & Wodak, S. J. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research* **37,** 825–831. ISSN: 1362-4962 (Feb. 2009).

308. Carvalho, S. B. *et al.* Intrinsically Disordered and Aggregation Prone Regions Underlie β-Aggregation in S100 Proteins. *PLoS ONE* **8** (ed Gasset, M.) e76629. ISSN: 1932-6203 (Oct. 2013).

309. Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21,** 3433–3434. ISSN: 13674803 (2005).

310. Pechmann, S., Levy, E. D., Tartaglia, G. G. & Vendruscolo, M. Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. *Proceedings of the National Academy of Sciences* **106,** 10159–10164. ISSN: 0027-8424 (June 2009).

311. Gabaldón, T., Rainey, D. & Huynen, M. A. Tracing the Evolution of a Large Protein Complex in the Eukaryotes, NADH:Ubiquinone Oxidoreductase (Complex I). *Journal of Molecular Biology* **348,** 857–870. ISSN: 00222836 (May 2005).

312. Huynen, M. A., Duarte, I. & Szklarczyk, R. Loss, replacement and gain of proteins at the origin of the mitochondria. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1827,** 224–231. ISSN: 00052728 (Feb. 2013).

313. Nasmyth, K. & Haering, C. H. Cohesin: Its Roles and Mechanisms. *Annual Review of Genetics* **43,** 525–558. ISSN: 0066-4197 (Dec. 2009).

314. Hirano, T. Condensin-Based Chromosome Organization from Bacteria to Vertebrates. *Cell* **164,** 847–857. ISSN: 0092-8674 (2016).

315. Kikuchi, S., Borek, D. M., Otwinowski, Z., Tomchick, D. R. & Yu, H. Crystal structure of the cohesin loader Scc2 and insight into cohesinopathy. *Proceedings of the National Academy of Sciences* **113,** 201611333. ISSN: 0027-8424 (2016).

316. Wood, A. J., Severson, A. F. & Meyer, B. J. Condensin and cohesin complexity: the expanding repertoire of functions. *Nature Reviews Genetics* **11,** 391–404. ISSN: 1471-0056 (June 2010).

317. Nasmyth, K. Disseminating the Genome: Joining, Resolving, and Separating Sister Chromatids During Mitosis and Meiosis. *Annual Review of Genetics* **35,** 673–745. ISSN: 0066-4197 (Dec. 2001).

318. Alipour, E. & Marko, J. F. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Research* **40,** 11202–11212. ISSN: 0305-1048 (Dec. 2012).

319. Goloborodko, A., Imakaev, M. V., Marko, J. F. & Mirny, L. Compaction and segregation of sister chromatids via active loop extrusion. *bioRxiv,* 1–20. ISSN: 2050-084X (2016).

320. Wang, X., Brandão, H. B., Le, T. B. K., Laub, M. T. & Rudner, D. Z. Bacillus subtilis SMC complexes juxtapose chromosome arms as they travel from origin to terminus. *Science* **355,** 524–527. ISSN: 0036-8075 (Feb. 2017).

321. Palecek, J. J. & Gruber, S. Kite Proteins: a Superfamily of SMC/Kleisin Partners Conserved Across Bacteria, Archaea, and Eukaryotes. *Structure* **23,** 2183–2190. ISSN: 09692126 (2015).

322. Ampatzidou, E., Irmisch, A., O'Connell, M. J. & Murray, J. M. Smc5/6 Is Required for Repair at Collapsed Replication Forks. *Molecular and Cellular Biology* **26,** 9387–9401. ISSN: 0270-7306 (Dec. 2006).

323. Farmer, S., San-Segundo, P. A. & Aragón, L. The Smc5–Smc6 Complex Is Required to Remove Chromosome Junctions in Meiosis. *PLoS ONE* **6** (ed Lichten, M.) e20948. ISSN: 1932-6203 (June 2011).

324. Uhlmann, F. SMC complexes: from DNA to chromosomes. *Nature Reviews Molecular Cell Biology* **17,** 399–412. ISSN: 1471-0072 (Apr. 2016).

325. Andrade, M. A. & Bork, P. HEAT repeats in the Huntington's disease protein. *Nature genetics* **11,** 115–6. ISSN: 1061-4036 (Oct. 1995).

326. Morin, P. J. beta-catenin signaling and cancer. *BioEssays : news and reviews in molecular, cellular and developmental biology* **21,** 1021–30. ISSN: 0265-9247 (Dec. 1999).

327. McMahon, H. T. & Mills, I. G. COP and clathrin-coated vesicle budding: different pathways, common approaches. *Current Opinion in Cell Biology* **16,** 379–391. ISSN: 09550674 (Aug. 2004).

328. Chook, Y. M. & Süel, K. E. Nuclear import by karyopherin-βs: Recognition and inhibition. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1813,** 1593–1606. ISSN: 01674889 (Sept. 2011).

329. Neuwald, A. F. & Hirano, T. HEAT Repeats Associated with Condensins, Cohesins, and Other Complexes Involved in Chromosome-Related Functions. *Genome Research* **10,** 1445–1452. ISSN: 10889051 (Oct. 2000).

330. Persi, E., Wolf, Y. I. & Koonin, E. V. Positive and strongly relaxed purifying selection drive the evolution of repeats in proteins. *Nature Communications* **7,** 13570. ISSN: 2041-1723 (Nov. 2016).

331. Andrade, M. a., Petosa, C., O'Donoghue, S. I., Müller, C. W. & Bork, P. Comparison of ARM and HEAT protein repeats. *Journal of molecular biology* **309,** 1–18. ISSN: 0022-2836 (May 2001).

332. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9,** 173–175. ISSN: 1548-7091 (2011).

333. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25,** 3389–402. ISSN: 0305-1048 (Sept. 1997).

334. Smith, L. *et al.* Candidate testis-determining gene, Maestro (Mro), encodes a novel HEAT repeat protein. *Developmental Dynamics* **227,** 600–607. ISSN: 1058-8388 (Aug. 2003).

335. Benjamini, Y. & Hochberg, Y. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing* 1995. doi:10.2307/2346101. arXiv: 95/57289 [0035-9246]. http://www.jstor.org/stable/2346101.

336. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics (Oxford, England)* **21,** 3448–9. ISSN: 1367-4803 (Aug. 2005).

337. Stephan, A. K., Kliszczak, M. & Morrison, C. G. The Nse2/Mms21 SUMO ligase of the Smc5/6 complex in the maintenance of genome stability. *FEBS Letters* **585,** 2907–2913. ISSN: 00145793 (Sept. 2011).

338. Alt, A. *et al.* Specialized interfaces of Smc5/6 control hinge stability and DNA association. *Nature Communications* **8,** 14011. ISSN: 2041-1723 (Jan. 2017).

339. Pebernard, S. *et al.* The Nse5-Nse6 dimer mediates DNA repair roles of the Smc5-Smc6 complex. *Mol Cell Biol* **26,** 1617–1630. ISSN: 0270-7306 (2006).

340. Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. E. Enhanced genome annotation using structural profiles in the program 3D-PSSM1. *Journal of Molecular Biology* **299,** 501–522. ISSN: 0022-2836 (2000).

341. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protocols* **10,** 845–858. ISSN: 1754-2189 (June 2015).

342. Biegert, A. & Soding, J. De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics* **24,** 807–814. ISSN: 1367-4803 (Mar. 2008).

343. Duan, X. *et al.* Architecture of the Smc5/6 complex of Saccharomyces cerevisiae reveals a unique interaction between the Nse5-6 subcomplex and the hinge regions of Smc5 and Smc6. *Journal of Biological Chemistry* **284,** 8507–8515. ISSN: 00219258 (2009).

344. Schlesner, M. *et al.* Identification of Archaea-specific chemotaxis proteins which interact with the flagellar apparatus. *BMC microbiology* **9,** 56. ISSN: 1471-2180 (Mar. 2009).

345. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521,** 173–179. ISSN: 0028-0836 (2015).

346. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541,** 353–358. ISSN: 0028-0836 (2017).

347. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic acids research* **39,** W29–37. ISSN: 1362-4962 (July 2011).

348. Tůmová, P., Uzlíková, M., Wanner, G. & Nohýnková, E. Structural organization of very small chromosomes: study on a single-celled evolutionary distant eukaryote Giardia intestinalis. *Chromosoma* **124,** 81–94. ISSN: 0009-5915 (Mar. 2015).

349. Eme, L., Trilles, A., Moreira, D. & Brochier-Armanet, C. The phylogenomic analysis of the anaphase promoting complex and its targets points to complex and modern-like control of the cell cycle in the last common ancestor of eukaryotes. *BMC Evolutionary Biology* **11,** 265. ISSN: 1471-2148 (Dec. 2011).

350. Ouyang, Z., Zheng, G., Tomchick, D. R., Luo, X. & Yu, H. Structural Basis and IP6 Requirement for Pds5-Dependent Cohesin Dynamics. *Molecular Cell,* 1–12. ISSN: 10972765 (2016).

351. Hara, K. *et al.* Structure of cohesin subcomplex pinpoints direct shugoshin-Wapl antagonism in centromeric cohesion. *Nature structural & molecular biology* **21,** 864–70. ISSN: 1545-9985 (2014).

352. Chao, W. C. H. *et al.* Structure of the cohesin loader Scc2. *Nature Communications* **8,** 13952. ISSN: 2041-1723 (Jan. 2017).

353. Lee, B.-G. *et al.* Crystal Structure of the Cohesin Gatekeeper Pds5 and in Complex with Kleisin Scc1. *Cell Reports,* 2108–2115. ISSN: 22111247 (2016).

354. Devos, D. P., Gräf, R. & Field, M. C. Evolution of the nucleus. *Current Opinion in Cell Biology* **28,** 8–15. ISSN: 09550674 (2014).

355. Baum, D. a. & Baum, B. An inside-out origin for the eukaryotic cell. *BMC biology* **12,** 76. ISSN: 1741-7007 (2014).

356. Case, R. B. The Bacterial Condensin MukBEF Compacts DNA into a Repetitive, Stable Structure. *Science* **305,** 222–227. ISSN: 0036-8075 (July 2004).

357. Niki, H. & Yano, K. In vitro topological loading of bacterial condensin MukB on DNA, preferentially single-stranded DNA rather than double-stranded DNA. *Scientific Reports* **6,** 29469. ISSN: 2045-2322 (Sept. 2016).

358. Kimura, M. On the probability of fixation of mutant genes in a population. *Genetics* **47,** 713–719. ISSN: 00166731 (1962).

359. Lynch, M., Bobay, L.-M., Catania, F., Gout, J.-F. & Rho, M. The Repatterning of Eukaryotic Genomes by Random Genetic Drift. *Annual Review of Genomics and Human Genetics* **12,** 347–366. ISSN: 1527-8204 (2011).

360. Mueller, S. *et al.* Protein degradation corrects for imbalanced subunit stoichiometry in OST complex assembly. *Molecular Biology of the Cell* **26,** 2596–2608. ISSN: 1059-1524 (2015).

361. Seidl, M. F. & Schultz, J. Evolutionary flexibility of protein complexes. *BMC Evolutionary Biology* **9,** 155. ISSN: 1471-2148 (2009).

362. Devos, D. *et al.* Simple fold composition and modular architecture of the nuclear pore complex. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 2172–7. ISSN: 0027-8424 (2006).

363. Field, M. C., Sali, A. & Rout, M. P. On a bender-BARs, ESCRTs, COPs, and finally getting your coat. *Journal of Cell Biology* **193,** 963–972. ISSN: 00219525 (2011).

364. Hurst, L. D., Pál, C. & Lercher, M. J. The evolutionary dynamics of eukaryotic gene order. *Nature reviews. Genetics* **5,** 299–310. ISSN: 1471-0056 (Apr. 2004).

365. Tan, K., Shlomi, T., Feizi, H., Ideker, T. & Sharan, R. Transcriptional regulation of protein complexes within and across species. *Proceedings of the National Academy of Sciences* **104,** 1283–1288. ISSN: 0027-8424 (Jan. 2007).

366. Tsai, H.-K., Su, C. P., Lu, M.-Y. J., Shih, C.-H. & Wang, D. Co-expression of adjacent genes in yeast cannot be simply attributed to shared regulatory system. en. *BMC Genomics* **8,** 352. ISSN: 14712164 (Oct. 2007).

367. Muhammad, D., Schmittling, S., Williams, C. & Long, T. A. More than meets the eye: Emergent properties of transcription factors networks in Arabidopsis. *Biochimica et biophysica acta* **1860,** 64–74. ISSN: 18749399 (Jan. 2017).

368. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489,** 83–90. ISSN: 0028-0836 (2012).

369. Graur, D. *et al.* On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of encode. *Genome Biology and Evolution* **5,** 578–590. ISSN: 17596653 (2013).

370. Graur, D. An Upper Limit on the Functional Fraction of the Human Genome. *Genome Biology and Evolution* **9,** 1880–1885. ISSN: 1759-6653 (July 2017).

371. Choudhary, C. *et al.* Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science (New York, N.Y.)* **325,** 834–40. ISSN: 1095-9203 (2009).

372. Caron, C., Boyault, C. & Khochbin, S. Regulatory cross-talk between lysine acetylation and ubiquitination: role in the control of protein stability. *BioEssays* **27,** 408–415. ISSN: 0265-9247 (Apr. 2005).

373. De Lichtenberg, U., Jensen, L. J., Brunak, S. & Bork, P. Dynamic complex formation during the yeast cell cycle. *Science* **307,** 724–7. ISSN: 1095-9203 (Feb. 2005).

374. Levy, E. D. PiQSi: Protein Quaternary Structure Investigation. *Structure* **15,** 1364–1367. ISSN: 09692126 (2007).

375. Mao, X. *et al.* DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic acids research* **42,** D654–9. ISSN: 1362-4962 (Jan. 2014).

376. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research* **41,** D808–D815. ISSN: 0305-1048 (Jan. 2013).

377. Okamura, Y. *et al.* COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic acids research,* 1–5. ISSN: 1362-4962 (Nov. 2014).

378. Dosztányi, Z., Csizmók, V., Tompa, P. & Simon, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of Molecular Biology* **347,** 827–839. ISSN: 00222836 (2005).

379. Soding, J. & Söding, J. Protein homology detection by HMM-HMM comparison. eng. *Bioinformatics* **21,** 951–960. ISSN: 13674803 (Apr. 2005).

380. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. eng. *Journal of molecular biology* **292,** 195–202. ISSN: 0022-2836 (Print) (Sept. 1999).

381. The PyMOL Molecular Graphics System, Version 1.8 Schrodinger, LLC.

382. Zhang, Y. & Skolnick, J. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* **33,** 2302–2309. ISSN: 03051048 (2005).

383. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. eng. *Journal of molecular biology* **215,** 403–410. ISSN: 0022-2836 (Print) (Oct. 1990).

384. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461. ISSN: 1367-4803 (Oct. 2010).

385. Katoh, K., Misawa, K., Kuma, K.-i. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* **30,** 3059–66. ISSN: 1362-4962 (July 2002).

386. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32,** 1792–1797. ISSN: 1362-4962 (Mar. 2004).

387. Ye, Y. *et al.* GLProbs: Aligning Multiple Sequences Adaptively. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **12,** 67–78. ISSN: 1545-5963 (Jan. 2015).

388. Collingridge, P. W. & Kelly, S. MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. *BMC bioinformatics* **13,** 117. ISSN: 1471-2105 (2012).

389. Pettersen, E. F. *et al.* UCSF Chimera–a visualization system for exploratory research and analysis. eng. *Journal of computational chemistry* **25,** 1605–1612. ISSN: 0192-8651 (Print) (Oct. 2004).